

# Marginal log-linear parameters for graphical Markov models

**Robin J. Evans**

Department of Statistics  
University of Washington  
rje42@stat.washington.edu

**Thomas S. Richardson**

Department of Statistics  
University of Washington  
tsr@stat.washington.edu

January 26, 2013

## Abstract

Marginal log-linear (MLL) models provide a flexible approach to multivariate discrete data. MLL parametrizations under linear constraints induce a wide variety of models, including models defined by conditional independences. We introduce a sub-class of MLL models which correspond to Acyclic Directed Mixed Graphs (ADMGs) under the usual global Markov property. We characterize for precisely which graphs the resulting parametrization is variation independent. The MLL approach provides the first description of ADMG models in terms of a minimal list of constraints. The parametrization is also easily adapted to sparse modelling techniques, which we illustrate using several examples of real data.

**Keywords:** acyclic directed mixed graph; discrete graphical model; marginal log-linear parameter; parsimonious modelling; variation independence.

## 1 Introduction

Models defined by conditional independence constraints are central to many methods in multivariate statistics, and in particular to graphical models (Darroch et al., 1980; Whittaker, 1990). In the case of discrete data, *marginal log-linear* (MLL) parameters can be used to parametrize a broad range of models, including some graphical classes and models for conditional independence (Rudas et al., 2010; Forcina et al., 2010). These parameters are defined by considering a sequence,  $M_1, M_2, \dots, M_k$ , of margins of the distribution which respects inclusion (i.e.  $M_i$  precedes  $M_j$  if  $M_i \subset M_j$ ), with each such sequence giving rise to a smooth parametrization of the saturated model. Useful sub-models can be induced by setting some of the parameters to zero, or more generally by restricting attention to a linear or affine subset of the parameter space.

The flexibility present in this scheme presents a challenge both in terms of interpreting the resulting model and performing model selection, for which a tractable search space is typically required. We describe a sub-class of marginal log-linear models corresponding to a class of graphs known as *acyclic directed mixed graphs* (ADMGs), which contain directed

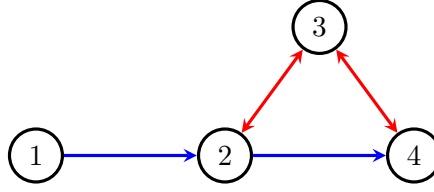


Figure 1: An acyclic directed mixed graph,  $\mathcal{G}_1$ .

( $\rightarrow$ ) and bidirected ( $\leftrightarrow$ ) edges, subject to the constraint that there are no cycles of directed edges; an example is given in Figure 1. The relationship between the MLL models and ADMGs is analogous to that between ordinary log-linear models and undirected graphs: log-linear models give a very rich class of models to choose from, since their number grows doubly-exponentially as the number of variables increases; undirected graphs provide a natural and more manageable subset of models with which to work (Darroch et al., 1980).

The patterns of independence described by ADMGs arise naturally in the context of generating processes in which not all variables are observed. To illustrate this, consider the randomized encouragement design carried out by McDonald et al. (1992) to investigate the effect of computer reminders for doctors on take-up of influenza vaccinations, and consequent morbidity in patients. The study involved 2,861 patients; here we focus on the following fields:

- (**Re**) *patient's doctor sent a card asking to **Remind** them about flu vaccine (randomized);*
- (**Va**) *patient **V**accinated against influenza;*
- (**Y**) *the endpoint: patient was not hospitalized with flu;*
- (**Ag**) ***A**ge of patient: 0 = '65 and under', 1 = 'over 65';*
- (**Co**) *patient has **C**hronic **O**bststructive Pulmonary Disease (**COPD**), as measured at base-line.*

The graphs in Figure 2 represent two possible data generating processes. Under both structures, whether or not a patient's doctor received a reminder note is independent of the baseline variables age (Ag) and COPD status (Co), as would be expected under randomization. Further the absence of an edge  $\text{Re} \rightarrow \text{Y}$  encodes the assumption that whether or not a reminder (Re) was received only influences the final outcome (Y) via whether or not a patient received a flu vaccination (Va). Both structures also assume that there are unobserved confounding factors between vaccination and COPD, and between COPD and the final outcome. However, the graph in Figure 2(b) supposes that there is no additional confounding between Va and Y. As a consequence the generating process given in (b) implies the additional restriction that  $\text{Re} \perp\!\!\!\perp \text{Y} \mid \text{Va}, \text{Ag}$ . (We make no assumptions about the state spaces of the variables H,  $\text{H}_1$  and  $\text{H}_2$ , since these factors are unobserved.)

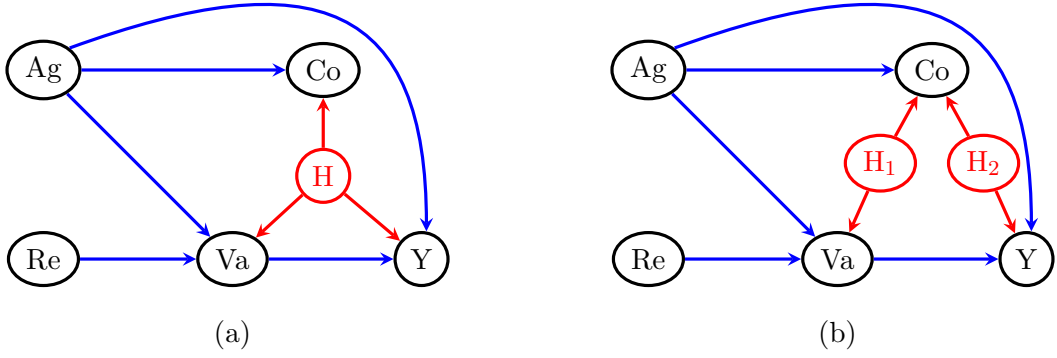


Figure 2: Two different generating processes for the flu vaccine encouragement design (red vertices are unobserved): both graphs imply  $\text{Re} \perp\!\!\!\perp \text{Ag}, \text{Co}$ ; however (b) also implies  $\text{Re} \perp\!\!\!\perp \text{Y} \mid \text{Va}, \text{Ag}$ .

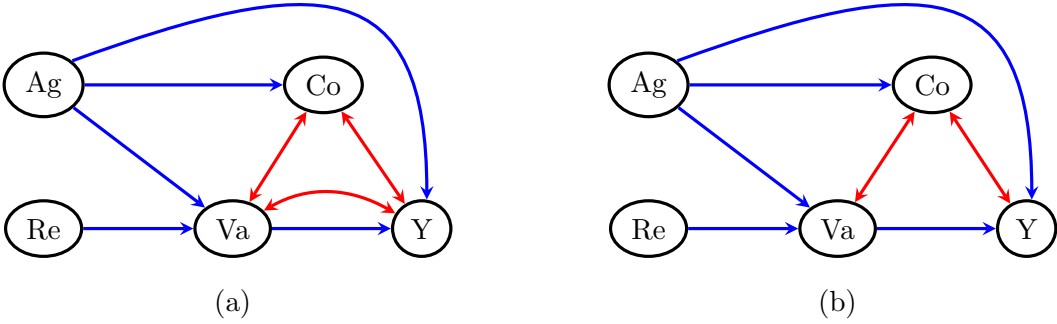


Figure 3: Two ADMGs representing the conditional independence restrictions on the observed margin implied by the corresponding graphs in Figure 2.

In Figure 3 we show the ADMGs corresponding to the generating processes in Figure 2. These graphs only contain observed variables, but by including bidirected edges ( $\leftrightarrow$ ) they encode the same observable conditional independence relations; see §3.1 for details.

All the work herein can easily be extended to graphs which also contain an undirected component, provided no undirected edge is adjacent to an arrowhead. This latter case is equivalent to the summary graphs of (Wermuth, 2011), and strictly includes all ancestral graphs (Richardson and Spirtes, 2002). Our approach may be seen as extending earlier work (Rudas et al., 2006, 2010; Forcina et al., 2010) which described the conditional independence structure of certain marginal log-linear models.

### 1.1 ADMG Models

Richardson (2003) described local and global Markov properties for ADMGs, while Richardson (2009) described a parametrization for discrete random variables via a collection of

conditional probabilities of the form  $P(X_H = 0 | X_T = x_T)$ . However, although Richardson’s parametrization is simple, it does not naturally lead to parsimonious sub-models. In addition, the parameters are subject to variation dependence constraints, in the sense that setting some parameters to particular values may restrict the valid range of other parameters; this makes maximum likelihood fitting, for example, more challenging (Evans and Richardson, 2010). To illustrate this point, consider the graph  $\mathcal{G}_1$  in Figure 1 as an example; it encodes the model under which  $X_1 \perp\!\!\!\perp X_3$  and  $X_4 \perp\!\!\!\perp X_1 | X_2$ . Richardson’s parametrization consists in this case (for binary random variables) of the probabilities

$$\begin{array}{lll} P(X_1 = 0) & P(X_2 = 0 | X_1 = x_1) & P(X_2 = 0, X_3 = 0 | X_1 = x_1) \\ P(X_3 = 0) & P(X_4 = 0 | X_2 = x_2) & P(X_3 = 0, X_4 = 0 | X_1 = x_1, X_2 = x_2) \end{array}$$

where  $x_1, x_2 \in \{0, 1\}$ . A disadvantage of this parametrization is that, for instance, the joint probabilities  $P(X_2 = 0, X_3 = 0 | X_1 = x_1)$  are bounded above by the marginal probabilities  $P(X_2 = 0 | X_1 = x_1)$ . Consequently, from the point of view of parameter interpretation, it makes little sense to consider the joint probabilities in isolation. For example, strong (conditional) correlation between  $X_2$  and  $X_3$  is present when the joint probability is large relative to the marginals.

However, replacing the joint probabilities  $P(X_2 = 0, X_3 = 0 | X_1 = x_1)$  with the conditional odds ratios

$$\frac{P(X_2 = 0, X_3 = 0 | X_1 = x_1) \cdot P(X_2 = 1, X_3 = 1 | X_1 = x_1)}{P(X_2 = 1, X_3 = 0 | X_1 = x_1) \cdot P(X_2 = 0, X_3 = 1 | X_1 = x_1)}, \quad x_1 \in \{0, 1\}$$

(and similarly for  $P(X_3 = 0, X_4 = 0 | X_1 = x_1, X_2 = x_2)$ ) yields a variation independent parametrization, the odds ratio measuring dependence without reference to marginal distributions. This means that if we wish to define a prior distribution over the univariate probabilities and the odds ratios, we may, if appropriate, simply use a product of univariate distributions; similarly, to fit a generalized linear model with these parameters as joint responses, we need only use simple univariate link functions. We will see that this approach to discrete parametrizations can be generalized using marginal log-linear parameters.

In Section 2 we introduce marginal log-linear (MLL) parameters and some of their properties, while Section 3 gives background theory about ADMGs and the parametrization of Richardson (2009). The development of MLL parameters for ADMG models is presented in Section 4, resulting in a parametrization we refer to as *ingenuous* (since it arises naturally, but ‘natural parametrization’ already has a particular meaning). We also show that this parametrization can always be embedded in a larger one corresponding to a complete graph and the saturated model, where some of the parameters in the bigger model are linearly constrained. In Section 5 we classify for which models the ingenuous parametrization is variation independent, since this can facilitate interpretation of the resulting coefficients. In Section 6 we discuss approaches to sparse modelling using MLLs in the context of several additional datasets and a simulation. Longer proofs are in Section 7.

## 2 Marginal Log-Linear Parameters

We consider collections of random variables  $(X_v)_{v \in V}$  with finite index set  $V$ , taking values in finite discrete probability spaces  $(\mathfrak{X}_v)_{v \in V}$  under a strictly positive probability measure  $P$ ; without loss of generality,  $\mathfrak{X}_v = \{0, 1, \dots, |\mathfrak{X}_v| - 1\}$ . For  $A \subseteq V$  we let  $\mathfrak{X}_A \equiv \times_{v \in A} (\mathfrak{X}_v)$ ,  $\mathfrak{X} \equiv \mathfrak{X}_V$  and similarly  $X_A \equiv (X_v)_{v \in A}$ ,  $X \equiv X_V$  and  $x_A \equiv (x_v)_{v \in A}$ ,  $x \equiv x_V$ . In addition  $\tilde{\mathfrak{X}}$  is the subset of  $\mathfrak{X}$  which does not contain the last possible element in any co-ordinate; that is  $\tilde{\mathfrak{X}}_v = \{0, 1, \dots, |\mathfrak{X}_v| - 2\}$ , and  $\tilde{\mathfrak{X}} = \times_{v \in V} (\tilde{\mathfrak{X}}_v)$ . We use  $p_A(x_A) \equiv P(X_A = x_A)$  and  $p_{A|B}(x_A | x_B) \equiv P(X_A = x_A | X_B = x_B)$ , for particular instantiations of  $x$ .

Following Bergsma and Rudas (2002), we define a general class of parameters on discrete distributions. The definition relies upon abstract collections of subsets, so it may be helpful to the reader to keep in mind that the sets  $M_i \in \mathbb{M}$  are margins, or subsets, of the distribution over  $V$ , and each set  $\mathbb{L}_i$  is a collection of effects in the margin  $M_i$ . A pair  $(L, M_i)$  corresponds to a log-linear interaction over the set  $L$ , within the margin  $M_i$ .

**Definition 2.1.** For  $L \subseteq M \subseteq V$ , the pair  $(L, M)$  is an ordered pair of subsets of  $V$ . Let  $\mathbb{P}$  be a collection of such pairs, and define

$$\mathbb{M} \equiv \{M \mid (L, M) \in \mathbb{P} \text{ for some } L\},$$

to be the collection of margins in  $\mathbb{P}$ . If  $\mathbb{M} = \{M_1, \dots, M_k\}$ , write

$$\mathbb{L}_i \equiv \{L \mid (L, M_i) \in \mathbb{P}\},$$

for the set of effects present in the margin  $M_i$ . We say that the collection  $\mathbb{P}$  is *hierarchical* if the ordering on  $\mathbb{M}$  may be chosen so that if  $i < j$ , then  $M_j \not\subseteq M_i$  and also  $L \in \mathbb{L}_j \Rightarrow L \not\subseteq M_i$ ; the second condition is equivalent to saying that each  $L$  is associated only with the first margin  $M$  of which it is a subset. We say the collection is *complete* if every non-empty subset of  $V$  is an element of precisely one set  $\mathbb{L}_i$ .

The term ‘hierarchical’ is used because each log-linear interaction is defined in the first possible margin in an ascending class, and ‘complete’ because all interactions are present. Some authors (Rudas et al., 2010; Lupparelli et al., 2009) consider only collections which are complete.

**Definition 2.2.** For each  $M \subseteq V$  and  $x_M \in \mathfrak{X}_M$ , define the functions  $\lambda_L^M(x_L)$  by the identity

$$\log p_M(x_M) \equiv \sum_{L \subseteq M} \lambda_L^M(x_L),$$

subject to the identifiability constraint that for every  $\emptyset \neq L \subseteq M$ ,  $x_L \in \mathfrak{X}_L$  and  $v \in L$ ,

$$\sum_{x_v \in \mathfrak{X}_v} \lambda_L^M(x_{L \setminus \{v\}}, x_v) = 0;$$

that is, the sum over the support of each variable is zero. We call  $\lambda_L^M(x_L)$  a *marginal log-linear parameter*.

Note that the constant  $\lambda_\emptyset^M$  is determined by the values of the other parameters and the fact that the probabilities  $p_M(x_M)$  sum to one. In the sequel we will always assume that  $L$  is non-empty.

The term ‘marginal log-linear parameter’ is coined by analogy with ordinary log-linear parameters, which correspond to the special case  $M = V$ . The following result provides an explicit expression for  $\lambda_L^M(x_L)$ .

**Lemma 2.3.** *For  $L \subseteq M \subseteq V$  and  $x_L \in \mathfrak{X}_L$  we have*

$$\lambda_L^M(x_L) = \frac{1}{|\mathfrak{X}_M|} \sum_{y_M \in \mathfrak{X}_M} \log p_M(y_M) \prod_{v \in L} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v=y_v\}} - 1). \quad (1)$$

This result is elementary, and its proof is omitted.

For a collection of ordered pairs of subsets  $\mathbb{P}$  (see Definition 2.1), we let

$$\tilde{\Lambda}(\mathbb{P}) = \{\lambda_L^M(x_L) \mid (L, M) \in \mathbb{P}, x_L \in \tilde{\mathfrak{X}}_L\}$$

be the collection of marginal log-linear parameters associated with  $\mathbb{P}$ . Note that we avoid the redundancy created by the identifiability constraint by only considering  $x_L \in \tilde{\mathfrak{X}}_L$ .

The definition of a marginal log-linear parameter we give is equivalent to the recursive one given in Bergsma and Rudas (2002); since both expositions are somewhat abstract, we invite the reader to consult the examples below for additional intuition. In particular note that for binary random variables, the product in (1) is always  $\pm 1$ . Bergsma and Rudas (2002, Theorem 2) show that any collection  $\tilde{\Lambda}(\mathbb{P})$  where  $\mathbb{P}$  is hierarchical and complete smoothly parametrizes the saturated model, that is, it parametrizes the set of all positive distributions on  $\mathfrak{X}$ .

The restriction that the parameters must sum to zero is required for identifiability, but different constraints can be used in its place. We might instead require that  $\lambda_L^M(x_L)$  be zero whenever any entry of  $x_L$  is zero (or some other selected value); this is seen in Marchetti and Lupporelli (2011), for example, and its use would not substantially affect any of the results in this paper.

## 2.1 Examples of Marginal Log-Linear Models

We will write  $\lambda_L^M$  to mean the collection  $\{\lambda_L^M(x_L) \mid x_L \in \mathfrak{X}_L\}$ ; the expression  $\lambda_L^M = 0$  denotes that we are setting all the parameters in this collection to 0.

**Example 2.4.** The classical log-linear parameters for a discrete distribution over a set of variables  $V$  are  $\{\lambda_L^V \mid L \subseteq V\}$ .

**Example 2.5.** Up to trivial transformations, the multivariate logistic parameters of Glonek and McCullagh (1995) are  $\{\lambda_L^L \mid L \subseteq V\}$ .

**Example 2.6.** Let  $V = \{1, 2, 3\}$  and assume all random variables are binary. Write  $P_{001} \equiv P(X_1 = 0, X_2 = 0, X_3 = 1)$ , and  $P_{1++} \equiv P(X_1 = 1)$ , etc. Then

$$\lambda_1^1(0) = \frac{1}{2} \log \frac{P_{0++}}{P_{1++}},$$

which, up to a multiplicative constant, is the logit of the probability of the event  $\{X_1 = 0\}$ . Also,

$$\lambda_1^{12}(0) = \frac{1}{4} \log \frac{P_{00+} P_{01+}}{P_{10+} P_{11+}} \quad \text{and} \quad \lambda_{12}^{12}(0, 0) = \frac{1}{4} \log \frac{P_{00+} P_{11+}}{P_{10+} P_{01+}},$$

the log odds product and log odds ratio between  $X_1$  and  $X_2$  respectively.

If instead  $X_1$  is ternary, we obtain

$$\lambda_1^1(0) = \frac{1}{3} \log \frac{P_{0++}^2}{P_{1++} P_{2++}},$$

$$\lambda_1^{12}(0) = \frac{1}{6} \log \frac{P_{00+}^2 P_{01+}}{P_{10+} P_{11+} P_{20+} P_{21+}} \quad \text{and} \quad \lambda_{12}^{12}(0, 0) = \frac{1}{6} \log \frac{P_{00+}^2 P_{11+} P_{21+}}{P_{10+} P_{20+} P_{01+}^2}.$$

Here  $\lambda_1^1(0)$  contrasts the probability  $P(X_1 = 0)$  with the geometric mean of the probabilities  $P(X_1 = 1)$  and  $P(X_1 = 2)$ . On the other hand, up to constants,  $\lambda_{12}^{12}(0, 0)$  is an average of the two log odds ratios

$$\log \frac{P_{00+} P_{21+}}{P_{20+} P_{01+}} \quad \log \frac{P_{00+} P_{11+}}{P_{10+} P_{01+}},$$

and so gives a contrast between  $P(X_1 = X_2 = 0)$  and other joint probabilities in a way which generalizes the binary log odds ratio and provides a measure of dependence; in particular note that  $\lambda_{12}^{12}(0, 0) = 0$  if  $X_1 \perp\!\!\!\perp X_2$ .

Here we have written, for example, 12 instead of  $\{1, 2\}$ ; similarly, for sets  $A$  and  $B$  we sometimes write  $AB$  for  $A \cup B$ , and  $aB$  for  $\{a\} \cup B$ .

## 2.2 Properties of Marginal Log-Linear Models

The next result relates marginal log-linear parameters to conditional independences; it is found as Lemma 1 in Rudas et al. (2010) and Equation (6) of Forcina et al. (2010).

**Lemma 2.7.** *For any disjoint sets  $A$ ,  $B$  and  $C$ , where  $C$  may be empty,  $A \perp\!\!\!\perp B \mid C$  if and only if*

$$\lambda_{A'B'C'}^{ABC} = 0 \quad \text{for every } \emptyset \neq A' \subseteq A, \quad \emptyset \neq B' \subseteq B, \quad C' \subseteq C.$$

The special case of  $C = \emptyset$  (giving marginal independence) is proved in the context of multivariate logistic parameters by Kauermann (1997).

**Example 2.8.** Take a complete and hierarchical parametrization of 3 variables,

$$\lambda_1^1 \quad \lambda_2^2 \quad \lambda_3^3 \quad \lambda_{12}^{12} \quad \lambda_{13}^{13} \quad \lambda_{23}^{123} \quad \lambda_{123}^{123}.$$

Then we can force  $X_1 \perp\!\!\!\perp X_3$  by setting  $\lambda_{13}^{13} = 0$ . Similarly  $X_2 \perp\!\!\!\perp X_3 \mid X_1$  corresponds to setting  $\lambda_{23}^{123} = \lambda_{123}^{123} = 0$ .

The following lemma shows that under conditional independence constraints, certain MLL parameters defined within different margins are equal.

**Lemma 2.9.** *Suppose that  $A \perp\!\!\!\perp B \mid C$ , and  $A$  is non-empty. Then for any  $D \subseteq C$ ,*

$$\lambda_{AD}^{ABC}(x_{AD}) = \lambda_{AD}^{AC}(x_{AD}), \quad \text{for each } x_{AD} \in \mathfrak{X}_{AD}.$$

The proof of this result is found in Section 7.1.

### 3 Acyclic Directed Mixed Graphs

We introduce basic graphical concepts used to describe the global Markov property and parametrization schemes.

**Definition 3.1.** A *directed mixed graph*  $\mathcal{G}$  consists of a set of vertices  $V$ , and both directed ( $\rightarrow$ ) and bidirected ( $\leftrightarrow$ ) edges. Edges of the same type and orientation may not be repeated, but there may be multiple edges of different types between a pair of vertices.

A *path* in  $\mathcal{G}$  is a sequence of adjacent edges, without repetition of a vertex; a path may be empty, or equivalently consist of only one vertex. The first and last vertices on a path are the *endpoints* (these are not distinct if the path is empty); other vertices on the path are *non-endpoints*. The graph  $\mathcal{G}_1$  in Figure 1, for example, contains the path  $1 \rightarrow 2 \rightarrow 4 \leftrightarrow 3$ , with endpoints 1 and 3, and non-endpoints 2 and 4. A *directed path* is one in which all the edges are directed ( $\rightarrow$ ) and are oriented in the same direction, whereas a *bidirected path* consists entirely of bidirected edges.

A directed cycle is a non-empty sequence of edges of the form  $v \rightarrow \cdots \rightarrow v$ . An *acyclic* directed mixed graph (ADMG) is one which contains no directed cycles.

**Definition 3.2.** For a graph  $\mathcal{G}$  and a subset of its vertices  $A \subseteq V$ , we denote by  $\mathcal{G}_A$  the *induced subgraph* formed by  $A$ ; that is, the graph containing the vertices  $A$ , and the edges in  $\mathcal{G}$  whose endpoints are both in  $A$ .

**Definition 3.3.** Let  $a$  and  $d$  be vertices in a mixed graph  $\mathcal{G}$ . If  $a = d$ , or there is a directed path from  $a$  to  $d$ , we say that  $a$  is an *ancestor* of  $d$ , and that  $d$  is a *descendant* of  $a$ . The sets of ancestors of  $d$  and descendants of  $a$  are denoted  $\text{an}_{\mathcal{G}}(d)$  and  $\text{de}_{\mathcal{G}}(a)$  respectively. If there is a directed path from  $a$  to  $d$  containing precisely one edge ( $a \rightarrow d$ ) then  $a$  is called a *parent* of  $d$ ; the set of vertices which are parents of  $d$  is written  $\text{pa}_{\mathcal{G}}(d)$ .



The *district* of  $a$ , denoted  $\text{dis}_{\mathcal{G}}(a)$ , is the set containing  $a$  and all vertices which are connected to  $a$  by a bidirected path. These definitions are applied disjunctively to sets of vertices, so that, for example,

$$\text{pa}_{\mathcal{G}}(W) \equiv \bigcup_{w \in W} \text{pa}_{\mathcal{G}}(w), \quad \text{dis}_{\mathcal{G}}(W) \equiv \bigcup_{w \in W} \text{dis}_{\mathcal{G}}(w).$$

A set of vertices  $A$  is *ancestral* if  $A = \text{ang}_{\mathcal{G}}(A)$ ; that is,  $A$  contains all its own ancestors.

**Example 3.4.** Consider the graph  $\mathcal{G}_1$  in Figure 1. We have

$$\text{ang}_{\mathcal{G}_1}(4) = \{1, 2, 4\} \quad \text{ang}_{\mathcal{G}_1}(\{2, 3\}) = \{1, 2, 3\}.$$

The district of 3 is the set  $\{2, 3, 4\}$ , and since 3 has no parents,  $\text{pa}_{\mathcal{G}_1}(3) = \emptyset$ .

Note that by the definitions of some authors, vertices are not their own ancestors (Lauritzen, 1996). The above notations may be shortened on induced subgraphs so that  $\text{pa}_A \equiv \text{pa}_{\mathcal{G}_A}$ , and similarly for other definitions. In some cases where the meaning is clear, we will dispense with the subscript altogether.

We use the now standard notation of Dawid (1979), and represent the statement ‘ $X$  is independent of  $Y$  given  $Z$  under a probability measure  $P$ ’, for random variables  $X$ ,  $Y$  and  $Z$ , by  $X \perp\!\!\!\perp Y \mid Z [P]$ . If  $P$  is unambiguous, this part is dropped, and if  $Z$  is empty we write simply  $X \perp\!\!\!\perp Y$ . Finally, we abuse notation in the usual way:  $v$  and  $X_v$  are used interchangeably as both a vertex and a random variable; likewise  $A$  denotes both a vertex set and  $X_A$ .

### 3.1 Global Markov Property for ADMGs

A Markov property associates a particular set of independence relations with a graph.

A non-endpoint vertex  $c$  on a path is a *collider* on the path if the edges preceding and succeeding  $c$  on the path have an arrowhead at  $c$ , for example  $\rightarrow c \leftarrow$  or  $\leftrightarrow c \leftarrow$ ; otherwise  $c$  is a *non-collider*. A path between vertices  $a$  and  $b$  in a mixed graph is said to be *blocked given a set  $C$*  if either

- (i) there is a non-collider on the path in  $C$ , or
- (ii) there is a collider on the path which is *not* in  $\text{ang}(C)$ .

If all paths from  $a$  to  $b$  are blocked by  $C$ , then  $a$  and  $b$  are said to be *m-separated given  $C$* . Sets  $A$  and  $B$  are said to be m-separated given  $C$  if every  $a \in A$  and every  $b \in B$  are m-separated given  $C$ . This naturally extends the d-separation criterion of Pearl (1988) to graphs with bidirected edges.

A probability measure  $P$  on  $\mathfrak{X}$  is said to satisfy the *global Markov property* for  $\mathcal{G}$  if for every triple of disjoint sets of vertices  $A$ ,  $B$  and  $C$ ,

$$A \text{ is m-separated from } B \text{ given } C \text{ in } \mathcal{G} \quad \implies \quad X_A \perp\!\!\!\perp X_B \mid X_C [P].$$

The *model* associated with an ADMG  $\mathcal{G}$  is simply the set of distributions that obey the global Markov property for  $\mathcal{G}$ .

**Proposition 3.5.** *If a path  $m$ -connects  $x$  and  $y$  given  $Z$  in  $\mathcal{G}$  then every vertex on the path is in  $\text{an}_{\mathcal{G}}(\{x, y\} \cup Z)$ .*

*Proof.* This follows from the definition of  $m$ -connection.  $\square$

**Example 3.6.** Consider the graph  $\mathcal{G}_1$  in Figure 1. There are two paths between the vertices 1 and 4,

$$\pi_1 : 1 \rightarrow 2 \rightarrow 4 \quad \text{and} \quad \pi_2 : 1 \rightarrow 2 \leftrightarrow 3 \leftrightarrow 4;$$

both are blocked by  $C = \{2\}$ .  $\pi_1$  is blocked because 2 is a non-collider on the path and is in  $C$ , while  $\pi_2$  is blocked because 3 is a collider on the path and is not in  $\text{an}_{\mathcal{G}_1}(C) = \{1, 2\}$ . Hence  $\{1\}$  and  $\{4\}$  are  $m$ -separated given  $\{2\}$  in  $\mathcal{G}_1$ .

One can similarly see that  $\{1\}$  and  $\{3\}$  are  $m$ -separated given  $C = \emptyset$ , and that no other  $m$ -separations hold for this graph. Thus a joint distribution  $P$  obeys the global Markov property for  $\mathcal{G}_1$  if and only if  $X_1 \perp\!\!\!\perp X_4 \mid X_2 [P]$  and  $X_1 \perp\!\!\!\perp X_3 [P]$ .

By similar arguments the independences associated with the ADMGs in Figure 2 may also be read off.

### 3.2 Existing Parametrization of ADMG models

This subsection defines the parameters of Richardson (2009) for multivariate discrete distributions satisfying the global Markov property for an ADMG.

**Definition 3.7.** Let  $\mathcal{G}$  be an ADMG with vertex set  $V$ . We say that a collection of vertices  $W \subseteq V$  is *barren* if for each  $v \in W$ , we have  $W \cap \text{deg}(v) = \{v\}$ ; in other words  $v$  has no non-trivial descendants in  $W$ . For an arbitrary set of vertices  $U$ , the maximal subset with no non-trivial descendants in  $U$  is denoted  $\text{barren}_{\mathcal{G}}(U)$ .

A *head* is a collection of vertices  $H$  which is connected by bidirected paths in  $\mathcal{G}_{\text{an}(H)}$  and is barren in  $\mathcal{G}$ . We write  $\mathcal{H}(\mathcal{G})$  for the collection of heads in  $\mathcal{G}$ . The *tail* of a head  $H$  is the set

$$\text{tail}_{\mathcal{G}}(H) \equiv \text{pa}_{\mathcal{G}}(\text{dis}_{\text{an}(H)}(H)) \cup (\text{dis}_{\text{an}(H)}(H) \setminus H).$$

Thus the tail of  $H$  is the set of vertices in  $V \setminus H$  connected to a vertex in  $H$  by a path on which every vertex is a collider and an ancestor of a vertex in  $H$ . We typically write  $T$  for a tail, provided it is clear which head it belongs to.

**Proposition 3.8.** *Let  $H$  be a head. Then (i)  $H = \text{barren}_{\mathcal{G}}(H \cup \text{tail}_{\mathcal{G}}(H))$ ; (ii)  $\text{tail}_{\mathcal{G}}(H) \subseteq \text{an}_{\mathcal{G}}(H)$ .*

*Proof.* Immediate from the respective definitions.  $\square$

Richardson (2009) shows that discrete distributions obeying the global Markov property for an ADMG  $\mathcal{G}$  are parametrized by the conditional probabilities:

$$\left\{ P(X_H = x_H \mid X_T = x_T) \mid H \in \mathcal{H}, T = \text{tail}_{\mathcal{G}}(H), x_H \in \tilde{\mathfrak{X}}_H, x_T \in \mathfrak{X}_T \right\}.$$

This is achieved via factorizations based on head-tail pairs; let  $\prec$  be the partial ordering on heads such that  $H_i \prec H_j$  if  $H_i \subset \text{an}_{\mathcal{G}}(H_j)$  and  $H_i \neq H_j$ . This is well defined, since otherwise  $\mathcal{G}$  would contain a directed cycle. Then let  $[\cdot]_{\mathcal{G}}$  be a function which partitions sets of vertices into heads by repeatedly removing heads which are maximal under  $\prec$ .

Then  $P$  satisfies the global Markov property for  $\mathcal{G}$  if and only if it obeys the factorizations

$$P(X_A = x_A) = \prod_{H \in [A]_{\mathcal{G}}} P(X_H = x_H \mid X_T = x_T) \quad (2)$$

for ancestral sets of vertices  $A$ ; see Richardson (2009) for details. In the case of a directed acyclic graph (DAG), this corresponds to the probability distribution of each vertex conditional on its parents.

**Example 3.9.** Consider again the ADMG  $\mathcal{G}_1$  in Figure 1; its head-tail pairs  $(H, T)$  are  $(1, \emptyset)$ ,  $(2, 1)$ ,  $(3, \emptyset)$ ,  $(23, 1)$ ,  $(4, 2)$  and  $(34, 12)$ . Multivariate binary distributions obeying the global Markov property with respect to  $\mathcal{G}_1$  are therefore parametrized by

$$\begin{aligned} p_1(0) \quad p_{2|1}(0 \mid x_1) \quad p_3(0) \quad p_{23|1}(0, 0, \mid x_1) \\ p_{4|2}(0 \mid x_2) \quad p_{34|12}(0, 0 \mid x_1, x_2), \end{aligned}$$

for  $x_1, x_2 \in \{0, 1\}$ , as mentioned in the Introduction.

### 3.3 Graphical Completions

Given a discrete model defined by a set of conditional independence constraints, it is natural to consider it as a sub-model of the saturated model, which contains all positive probability distributions. In a setting where the model is graphical, it becomes equally natural to think of the graph as a subgraph of a complete graph, by which we mean a graph containing at least one edge between every pair of vertices. We can obtain a complete graph from an incomplete one by inserting edges between each pair of vertices which lack one, but this leaves a choice of edge type and orientation. These choices may affect how much of the structure and spirit of the original graph is retained; we will require that a completion preserves the heads of the original graph, which helps to preserve the structure of the parametrization.

**Definition 3.10.** Given an ADMG  $\mathcal{G}$  and a supergraph  $\bar{\mathcal{G}}$ , we call  $\bar{\mathcal{G}}$  a *head-preserving completion* of  $\mathcal{G}$  if  $\bar{\mathcal{G}}$  is complete, and  $\mathcal{H}(\mathcal{G}) \subseteq \mathcal{H}(\bar{\mathcal{G}})$ .

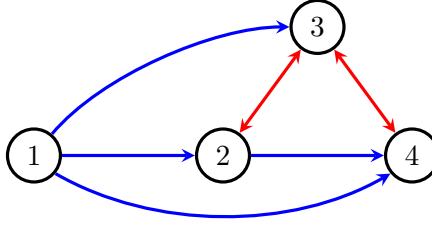


Figure 4: A head-preserving completion,  $\bar{\mathcal{G}}_1$ , of the ADMG in Figure 1.

It is easy to see that a head-preserving completion always exists; for example, if we add in a bidirected edge between every pair of vertices which are not joined by an edge, then it is clear that barren sets in  $\mathcal{G}$  will remain barren in  $\bar{\mathcal{G}}$ , and bidirected connected sets in  $\mathcal{G}$  will remain bidirected connected in  $\bar{\mathcal{G}}$ .

Note that it is not necessary for every pair of vertices to be joined by an edge in order for a graph to represent the saturated model, however we will require this for our completions.

**Example 3.11.** Figure 4 shows a head-preserving completion of the ADMG in Figure 1.

**Proposition 3.12.** *If  $\bar{\mathcal{G}}$  is a head-preserving completion of  $\mathcal{G}$  then  $\text{an}_{\mathcal{G}}(v) \subseteq \text{an}_{\bar{\mathcal{G}}}(v)$ . In particular, if a set  $A$  is ancestral in  $\bar{\mathcal{G}}$  then  $A$  is also ancestral in  $\mathcal{G}$ .*

*Proof.* This follows because  $\mathcal{G}$  contains a subset of the edges in  $\bar{\mathcal{G}}$ . □

## 4 Ingenuous Parametrization of an ADMG model

We now use the marginal log-linear parameters defined in Section 2 to parametrize the ADMG models discussed in Section 3.

**Definition 4.1.** Consider an ADMG  $\mathcal{G}$  with head-tail pairs  $(H_i, T_i)$  over some index  $i$ , and let  $M_i = H_i \cup T_i$ . Further, let  $\mathbb{L}_i = \{A \mid H_i \subseteq A \subseteq H_i \cup T_i\}$ . This collection of margins and associated effects is the *ingenuous* parametrization of  $\mathcal{G}$ , denoted  $\mathbb{P}^{\text{ing}}(\mathcal{G})$ .

**Example 4.2.** We return again to the ADMG  $\mathcal{G}_1$  in Figure 1; the head-tail pairs are  $(1, \emptyset)$ ,  $(2, 1)$ ,  $(3, \emptyset)$ ,  $(23, 1)$ ,  $(4, 2)$  and  $(34, 12)$ , meaning that the ingenuous parametrization is given by the following margins and effects:

$M$	$\mathbb{L}$
1	1
12	2, 12
3	3
123	23, 123
24	4, 24
1234	34, 134, 234, 1234.

Note that the ordering of the margins given here is hierarchical; in order to use most of the results of Bergsma and Rudas (2002), we need to confirm that the definition above always leads to a hierarchical parametrization, which is shown by the following result.

**Lemma 4.3.** *For any ADMG  $\mathcal{G}$ , there is an ordering on the margins  $M_i$  of the ingenuous parametrization  $\mathbb{P}^{\text{ing}}(\mathcal{G})$  which is hierarchical.*

*Proof.* Firstly we show that for distinct heads  $H_i$  and  $H_j$ , the collections  $\mathbb{L}_i$  and  $\mathbb{L}_j$  are disjoint. To see this, assume for a contradiction that there exists  $A$  such that  $H_i \subseteq A \subseteq H_i \cup T_i$  and  $H_j \subseteq A \subseteq H_j \cup T_j$ . Since  $H_i \neq H_j$ , assume without loss of generality that there exists  $v \in H_i \cap H_j^c \subseteq A$ .

Then  $v \in H_j \cup T_j$  implies that  $v \in T_j$ , and thus there is a directed path from  $v$  to some  $w \in H_j$ . Now,  $w \notin H_i$ , since  $v, w \in H_i$  would imply that  $H_i$  is not barren. But if  $w \in H_j \cap H_i^c$ , then by the same argument as above we can find a directed path from  $w$  to some  $x \in H_i$ . Then  $v \rightarrow \dots \rightarrow w \rightarrow \dots \rightarrow x$  is a directed path between elements of  $H_i$ , which is a contradiction. Thus  $\mathbb{L}_i$  and  $\mathbb{L}_j$  are disjoint.

Now, consider the partial ordering  $\prec$  of heads defined in Section 3.2:  $H_i \prec H_j$  whenever  $H_i \subset \text{ang}(H_j)$  and  $H_i \neq H_j$ . Any total ordering which respects this partial ordering is hierarchical, because each set  $A \in \mathbb{L}_i$  is a subset of the ancestors of  $H_i$ .  $\square$

We proceed to show that the ingenuous parameters for an ADMG  $\mathcal{G}$  characterize the set of distributions which obey the global Markov property with respect to  $\mathcal{G}$ .

**Lemma 4.4.** *For any sets  $M$  and  $L \subseteq M$ , the collection of MLL parameters*

$$\{\lambda_A^M(x_A) \mid L \subseteq A \subseteq M, x_M \in \tilde{\mathfrak{X}}_M\},$$

*together with the  $(|L| - 1)$ -dimensional marginal distributions of  $X_L$  conditional on  $X_{M \setminus L}$ , smoothly parametrizes the distribution of  $X_L$  conditional on  $X_{M \setminus L}$ .*

A proof is given in Section 7.2.

We now come to the main result of this section.

**Theorem 4.5.** *The ingenuous parametrization  $\tilde{\Lambda}(\mathbb{P}^{\text{ing}}(\mathcal{G}))$  of an ADMG  $\mathcal{G}$  parametrizes precisely those distributions  $P$  obeying the global Markov property with respect to  $\mathcal{G}$ .*

*Proof.* We proceed by induction. Again we use the partial ordering  $\prec$  on heads from Section 3.2. For the base case, we know that singleton heads  $\{h\}$  with empty tails are parametrized by the logits  $\lambda_h^h$ .

Now, suppose that we wish to find the distribution of a head  $H$  conditional on its tail  $T$ . Assume that we have the distribution of all heads  $H'$  which precede  $H$ , conditional on their respective tails; we claim this is sufficient to give the  $(|H| - 1)$ -dimensional marginal distributions of  $H$  conditional on  $T$ .

Let  $v \in H$ , and let  $C = H \setminus \{v\}$  be a  $(|H| - 1)$ -dimensional marginal of interest. The set  $A = \text{an}_{\mathcal{G}}(H) \setminus \{v\}$  is ancestral, since  $v$  cannot have (non-trivial) descendants in  $\text{an}_{\mathcal{G}}(H)$ ; in particular  $C \cup T \subseteq A$ . Theorem 4 of Richardson (2009) states that the factorization in equation (2) holds for every ancestral set, so

$$p_A(x_A) = \prod_{\substack{H' \in [A]_{\mathcal{G}} \\ T' = \text{tail}(H')}} p_{H'|T'}(x_{H'} | x_{T'}).$$

But all the probabilities in the product are known by our induction hypothesis, and the marginal distribution of  $C$  conditional on  $T$  is given by the distribution of  $A$ .

The ingenuous parametrization, by definition, contains  $\lambda_A^{H \cup T}$  for  $H \subseteq A \subseteq H \cup T$ , and thus the result follows from Lemma 4.4.  $\square$

**Example 4.6.** Returning to our running example, the graph  $\mathcal{G}_1$  in Figure 1 corresponds to the model

$$\left\{ P \mid X_1 \perp\!\!\!\perp X_4 \mid X_2 [P] \text{ and } X_1 \perp\!\!\!\perp X_3 [P] \right\}.$$

Theorem 4.5 tells us that this collection of distributions is precisely characterized by the ingenuous parameters for  $\mathcal{G}_1$ ,

$$\begin{array}{cccccc} \lambda_1^1 & \lambda_2^{12} & \lambda_{12}^{12} & \lambda_3^3 & \lambda_{23}^{123} & \lambda_{123}^{123} \\ \lambda_4^{24} & \lambda_{24}^{24} & \lambda_{34}^{1234} & \lambda_{134}^{1234} & \lambda_{234}^{1234} & \lambda_{1234}^{1234}. \end{array}$$

#### 4.1 Constraint-Based Model Description

The results above show that the ingenuous parameters for an ADMG  $\mathcal{G}$ , like Richardson's parameters, provide precisely the information required to reconstruct a distribution obeying the global Markov property for  $\mathcal{G}$ . However, it is difficult to use this parametrization in practice unless we can evaluate the likelihood, which requires us to make explicit the map which we have implicitly defined from the ingenuous parameters to the joint probability distribution under the model. For example, for the parameters in Richardson (2009) there is an explicit map from the parameters back to the joint distribution using a generalization of Möbius inversion. This was used by Evans and Richardson (2010) to fit these models via maximum likelihood. In contrast, the map from ingenuous parameters to the joint distribution cannot be written in closed form.

An alternative approach is to consider the ingenuous parametrization as part of a larger, complete parametrization of the saturated model, such that the additional parameters are constrained to be zero under the sub-model defined by  $\mathcal{G}$ . This enables us to fit the model using Lagrange-type algorithms, as in Evans and Forcina (2011).

**Theorem 4.7.** *Let  $\mathcal{G}$  be an ADMG, and  $\bar{\mathcal{G}}$  a head-preserving completion of  $\mathcal{G}$ . The ingenuous parametrization of  $\mathcal{G}$  corresponds to setting*

$$\lambda_L^M = 0$$

*for  $(L, M) \in \mathbb{P}^{\text{ing}}(\bar{\mathcal{G}})$  whenever  $L$  does not appear as an effect in  $\mathbb{P}^{\text{ing}}(\mathcal{G})$ . In particular, these constraints define the set of distributions which satisfy the global Markov property with respect to  $\mathcal{G}$ .*

The proof of this result is found in Section 7.3

**Example 4.8.** Consider again the ADMG  $\mathcal{G}_1$  in Figure 1; a possible head-preserving completion  $\bar{\mathcal{G}}_1$  (shown in Figure 4) is obtained by adding the edges  $1 \rightarrow 3$  and  $1 \rightarrow 4$ . The ingenuous parametrization for  $\bar{\mathcal{G}}_1$  is

$M$	$\mathbb{L}$
1	1
2	2, 12
13	3, 13
123	23, 123
124	4, 14, 24, 124
1234	34, 134, 234, 1234.

The effects found in  $\mathbb{P}^{\text{ing}}(\bar{\mathcal{G}}_1)$  but not in  $\mathbb{P}^{\text{ing}}(\mathcal{G}_1)$  are 13, 14, and 124, and indeed the sub-model defined by  $\mathcal{G}_1$  corresponds to setting

$$\lambda_{13}^{13} = \lambda_{14}^{124} = \lambda_{124}^{124} = 0;$$

under this model the following equalities hold by Lemma 2.9:

$$\lambda_4^{124} = \lambda_4^{24} \qquad \lambda_{24}^{124} = \lambda_{24}^{24}.$$

Removing the zero parameters in  $\mathbb{P}^{\text{ing}}(\bar{\mathcal{G}}_1)$  and renaming two others according to the above equations returns us to the ingenuous parametrization of  $\mathcal{G}_1$ .

Theorem 4.7 shows that we can fit the model defined by  $\mathcal{G}_1$  by maximum likelihood simply by maximizing the log-likelihood subject to  $\lambda_{13}^{13} = \lambda_{14}^{124} = \lambda_{124}^{124} = 0$ . In particular, this approach always provides a list of independent constraints which characterize the model.

An obvious question which arises is whether *any* completion of a graph will lead to a complete parametrization with the property of Theorem 4.7. We can obtain a counterexample by considering the complete graph  $\tilde{\mathcal{G}}_1$  in Figure 5, which has ingenuous parametrization

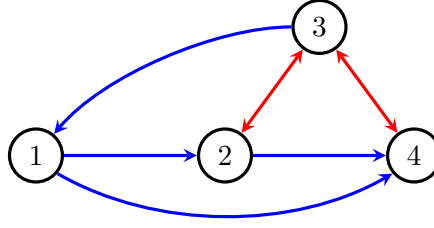


Figure 5: A complete ADMG,  $\tilde{\mathcal{G}}_1$ , of which  $\mathcal{G}_1$  is a subgraph, but whose ingenuous parametrization does not contain the model described by  $\mathcal{G}_1$  as a linear sub-space because the associated completion is not head-preserving.

$M$	$\mathbb{L}$
3	3
13	1, 13
123	2, 12, 23, 123
1234	4, 14, 24, 124, 34, 134, 234, 1234.

The graph  $\mathcal{G}_1$  in Figure 1 is a subgraph of  $\tilde{\mathcal{G}}_1$ , and corresponds to the model obtained by setting  $\lambda_{13}^{13} = \lambda_{14}^{124} = \lambda_{124}^{124} = 0$ ; however, these last two parameters do not appear in the ingenuous parametrization of  $\tilde{\mathcal{G}}_1$ , and so there is no way to enforce the sub-model as a linear constraint.

$\tilde{\mathcal{G}}_1$  is, of course, not head-preserving. Such completions may still lead to parametrizations which satisfy the property of Theorem 4.7: for example, if the edge  $1 \rightarrow 3$  is added to the graph in Figure 6(a), this destroys the head  $\{1, 2, 3\}$ , but the sub-model corresponds to  $\lambda_{13}^{13} = 0$ , which is a parameter in the complete graph.

## 4.2 Relationship To Prior Work

Rudas et al. (2010) parametrize chain graph models of multivariate regression type, also known as type IV chain graph models, using marginal log-linear parameters. Type IV chain graph models are a special case of ADMG models, in the sense that by replacing the undirected edges in a type IV chain graph with bidirected edges, the global Markov property on the resulting ADMG is equivalent to the Markov property for the chain graph (see Drton, 2009). The graphs in Figure 6 are examples of Type IV models. However, there are models in the class of ADMGs which do not correspond to any chain graph, such as the one described by  $\mathcal{G}_1$  in Figure 1.

The parametrization of Rudas et al. (2010) uses different choices of margins to the ingenuous parametrization, though their parameters can be shown to be equal to the parameters considered here under the global Markov property, using Lemma 2.9. Thus the variation dependence properties of that parametrization are identical to those of the ingenuous parametrization (see next section). Forcina et al. (2010) provide an algorithm which gives a range of ‘admissible’ margins in which collections of conditional independence



constraints may be defined.

Marchetti and Lupporelli (2011) also parametrize type IV chain graph models in a similar manner to Rudas et al. (2010), in that case using multivariate logistic contrasts.

## 5 Variation Independence

As discussed in the introduction, the interpretation of parameters and the construction of prior distributions is simpler when parameters are variation independent.

**Definition 5.1.** Let  $\theta_i$ , for  $i = 1, \dots, k$  be a collection of parameters such that  $\theta_i$  takes all values in the set  $\Theta_i$ . We say that the vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$  is *variation independent* if  $\boldsymbol{\theta}$  can take every value in the set  $\Theta_1 \times \dots \times \Theta_k$ .

Bergsma and Rudas (2002) characterize precisely which hierarchical and complete parametrizations are variation independent, using a notion they call ordered decomposability. We now do this for ingenuous parametrizations.

**Definition 5.2.** A collection of sets  $\mathbb{M} = \{M_1, \dots, M_k\}$  is *incomparable* if  $M_i \not\subseteq M_j$  for every  $i \neq j$ .

A collection  $\mathbb{M}$  of incomparable subsets of  $V$  is *decomposable* if it has at most two elements, or there is an ordering  $M_1, \dots, M_k$  on the elements of  $\mathbb{M}$  wherein for each  $i = 3, \dots, k$ , there exists  $j_i < i$  such that

$$\left( \bigcup_{l=1}^{i-1} M_l \right) \cap M_i = M_{j_i} \cap M_i.$$

This is also known as the *running intersection property*.

A collection  $\mathbb{M}$  of (possibly comparable) subsets is *ordered decomposable* if it has at most two elements, or there is an ordering  $M_1, \dots, M_k$  such that  $M_i \not\subseteq M_j$  for  $i > j$ , and for each  $i = 3, \dots, k$ , the inclusion maximal elements of  $\{M_1, \dots, M_i\}$  form a decomposable collection. We say that a collection  $\mathbb{P}$  of parameters is ordered decomposable if there is an ordering on the margins  $\mathbb{M}$  which is both hierarchical and ordered decomposable.

The following example is found in Bergsma and Rudas (2002).

**Example 5.3.** Let  $\mathbb{M} = \{12, 13, 23, 123\}$ . In order to have a hierarchical ordering of these margins it is clear that the set 123 must come last, but there is no way to order the collection of inclusion maximal margins  $\{12, 13, 23\}$  such that it has the running intersection property. Thus  $\mathbb{M}$  is not ordered decomposable.

The next result links variation independence to ordered decomposability.

**Theorem 5.4** (Bergsma and Rudas (2002), Theorem 4). *Let  $\mathbb{P}$  be a parametrization which is hierarchical and complete. Then the parameters  $\tilde{\Lambda}(\mathbb{P})$  are variation independent if and only if  $\mathbb{P}$  is ordered decomposable.*

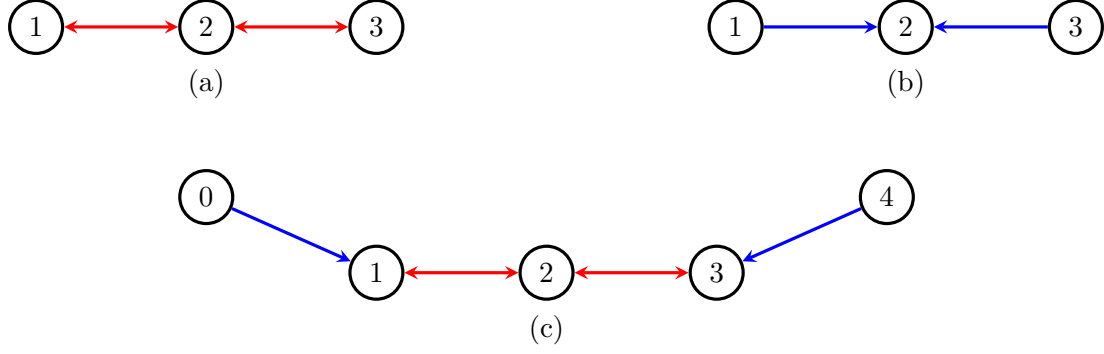


Figure 6: (a) a graph with a variation dependent ingenious parametrization; (b) a Markov equivalent graph to (a) with a variation independent ingenious parametrization; (c) a graph with no variation independent MLL parametrization.

As previously noted, the ingenious parametrization is not complete in general, and so we cannot apply the above result directly to characterize its variation dependence. However, by constructing complete parametrizations of which the ingenious parametrizations are linear sub-models, we can obtain the following.

**Theorem 5.5.** *The ingenious parametrization for an ADMG  $\mathcal{G}$  is variation independent if and only if  $\mathcal{G}$  contains no heads of size greater than or equal to 3.*

The proof of this result is found in Section 7.4.

**Example 5.6.** The graph  $\mathcal{G}_1$  in Figure 1 has maximum head size 2, and therefore the associated ingenious parametrization is variation independent.

Likewise the graphs in Figure 3(a) and (b) contain no heads of size greater than 2, so that the resulting ingenious parameters are variation independent. Note that this was not true of the parameters given by Richardson (2009).

**Example 5.7.** The bidirected 3-chain shown in Figure 6(a) has the head 123, and therefore its ingenious parametrization is variation dependent. This can easily be seen directly: in the binary case, for example, if the parameters  $\lambda_{12}^{12}(0)$  and  $\lambda_{23}^{23}(0)$  are chosen to be very large, this induces very strong dependence between the variables  $X_1$  and  $X_2$ , and between  $X_2$  and  $X_3$  respectively. If these correlations are chosen to be too large, then it is impossible for  $X_1$  and  $X_3$  to be marginally independent, which is implied by the graph.

Observe that we could use the Markov equivalent graph in Figure 6(b), which has no heads of size 3, and thus obtain a variation independent parametrization of the same model. However, if we add incident arrows as shown in Figure 6(c), we obtain a graph where such a trick is not possible. In fact this third graph has no variation independent parametrization in the Bergsma and Rudas framework, since it requires  $\lambda_{0124}^{0124} = \lambda_{0134}^{0134} = \lambda_{0234}^{0234} = 0$ , and these margins cannot be ordered in a way which satisfies the running intersection property

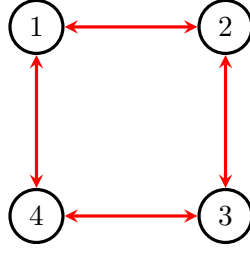


Figure 7: A bidirected 4-cycle.

(see Example 5.3).

In general, it would be sensible for a statistician concerned about variation dependence to choose a graph from the Markov equivalence class created by their model which has the smallest possible maximum head size. This could be achieved by reducing the number of bidirected edges in the graph, where possible; see, for example, Ali et al. (2005) and Drton and Richardson (2008b) for algorithms for finding the graph with the minimal number of arrowheads in a given Markov equivalence class.

**Example 5.8.** The bidirected 4-cycle, shown in Figure 7, contains a head of size 4, and so its ingenuous parametrization is variation dependent. However, there is a marginal log-linear parametrization of this model which is ordered decomposable, and therefore variation independent. The 4-cycle is precisely the model with  $X_1 \perp\!\!\!\perp X_3$  and  $X_2 \perp\!\!\!\perp X_4$ . Set  $\mathbb{M} = \{13, 24, 1234\}$ , with

$$\begin{aligned}\mathbb{L}_1 &= \{1, 3, 13\} \\ \mathbb{L}_2 &= \{2, 4, 24\} \\ \mathbb{L}_3 &= \mathcal{P}(\{1, 2, 3, 4\}) \setminus (\mathbb{L}_1 \cup \mathbb{L}_2);\end{aligned}$$

here  $\mathcal{P}(A)$  denotes the power set of  $A$ . This gives a hierarchical, complete and ordered decomposable parametrization, so the parameters are variation independent. The 4-cycle corresponds exactly to setting  $\lambda_{13}^{13} = \lambda_{24}^{24} = 0$ , and it follows that the remaining parameters are still variation independent under this constraint.

This approach to parametrization, which considers disconnected sets, is discussed in detail by Lupporelli et al. (2009). It produces a variation independent parametrization for graphs where the disconnected sets do not overlap, and may well be preferable to the ingenuous parametrization in these cases. In sparser graphs however, it does not seem as useful; as mentioned above, some graphs have no variation independent MLL parametrization.

## 6 Parsimonious Modelling with Marginal Log-Linear Parameters

The number of parameters in a model associated with a sparse graph containing bidirected edges can, in certain cases, be relatively large. In a purely bidirected graph, the parameter count depends upon the number of connected sets of vertices; in the case of a chain of bidirected edges such as that shown in Figure 11(a), this means that the number of parameters grows quadratically in the length of the chain.

The parametrization of Richardson (2009), and its special case for purely bidirected graphs (see Drton and Richardson, 2008a) does not present us with any obvious method of reducing the parameter count whilst preserving the conditional independence structure. In contrast, there are well established methods for sparse modelling with other classes of graphical models. In the case of an undirected graph with binary random variables, restricting to one parameter for each vertex and each edge leads to a Boltzmann Machine (Ackley et al., 1985). Rudas et al. (2006) use marginal log-linear parameters to provide a sparse parametrization of a DAG model, again restricting to one parameter for each vertex and edge.

As we will see from the following examples, the ingenuous parametrization allows us to fit graphical models with a large number of parameters, and then remove higher-order interactions to obtain a more parsimonious model whilst preserving the conditional independence structure of the original graph.

### 6.1 Flu Vaccination Data Revisited

We first return to the McDonald et al. (1992) study considered in the Introduction. All variables are binary, and (excepting Age) are coded as 0 = false, 1 = true; we add constraints to our model sequentially, recording the results in the analysis of deviance Table 1. The ADMG in Figure 3(a) represents the constraint  $\text{Ag}, \text{Co} \perp\!\!\!\perp \text{Re}$ ; it fits well, having a deviance of 2.54 on 3 degrees of freedom. The smaller model for 3(b) encodes

$$\text{Ag}, \text{Co} \perp\!\!\!\perp \text{Re} \qquad \text{Y} \perp\!\!\!\perp \text{Re} \mid \text{Va}, \text{Ag};$$

note that these precise independences cannot be represented by a DAG or chain graph (of any of the types considered by Drton (2009)). It also fits well (deviance 7.66 on 7 d.f.), so we may prefer it on the grounds of simplicity.

The ingenuous parametrization in this case contains some higher order effects, including the 5-way interaction between all variables. Setting  $\lambda_L^M = 0$  for  $|L| \geq 4$  removes five parameters whilst increasing the deviance by only 2.22; removing the effects of size 3 adds a further 8.39 to the deviance whilst removing seven more parameters. The resulting model has a total deviance of 18.28 on 19 degrees of freedom, representing a good fit compared to the saturated model (likelihood ratio test  $p = 0.49$ ).

Constraint	Figure	Add. Dev.	d.f.	Total Dev.
$Ag, Co \perp Re$	3(a)	2.54	3	2.54
$Y \perp Re \mid Va, Ag$	3(b)	5.11	7	7.66
no 4- and 5-way params		2.22	12	9.88
no 3-way params		8.39	19	18.28

Table 1: Analysis of deviance table of models considered for influenza data. Constraints are added sequentially from top to bottom; the last three columns give the additional deviance for the constraint, the total degrees of freedom and the total deviance of the models respectively.

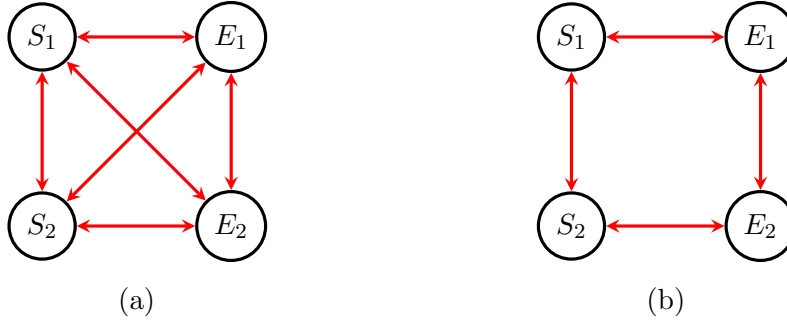


Figure 8: Graphs for the twins data for models corresponding to (a) a common gene and (b) separate genes affecting the prevalence of frozen shoulder and tennis elbow.

## 6.2 Incorporating Symmetry: Twins Data

Hakim et al. (2003) investigate genetic effects on the presence or absence of two soft tissue disorders, frozen shoulder and tennis elbow, based on a study in pairs of monozygotic and dizygotic twins; the data are reproduced in Ekholm et al. (2012). We have count data for a 5-way contingency table over the variables  $S_i$  and  $E_i$ , indicators of whether twin  $i$  in the pair suffers from frozen shoulder and tennis elbow respectively,  $i \in \{1, 2\}$ , and  $T$ , an indicator of whether the pair are monozygotic or dizygotic twins. There are a total of 866 observations for monozygotic pairs, and 963 for dizygotic pairs; twin 1 corresponds to the twin who was born first.

We first fitted the model  $T \perp (S_1, S_2, E_1, E_2)$  to test whether the zygosity of the twins has any effect on the other variables; we obtained a deviance of 16.4 on 15 degrees of freedom, suggesting that there is no evidence that  $T$  is related to the other variables. Note that this contradicts the conclusions of Ekholm et al. (2012), but they use additional assumptions to obtain more powerful tests.

Collapsing to a 4-way table over  $(S_1, S_2, E_1, E_2)$ , we consider the complete bidirected model in Figure 8(a). A further simplifying assumption is to impose symmetry between the twins in each pair, on the basis that we do not expect any association between the prevalence of the disorders and which twin was born first. Using the ingenuous parametrization

for the graph in Figure 8(a), which is itself symmetric with respect to the individual twins, this amounts to six independent linear constraints, and gives a deviance of 0.59 compared to the saturated model on four variables; there is therefore no evidence to reject symmetry.

Now, a hypothesis of interest is whether a common gene is responsible for the increased risk of the two disorders, or the genetic effects are separate and independent. In the latter case we would expect the data to be explained by the model encoded by the graph in Figure 8(b), and therefore to observe the marginal independences  $E_1 \perp\!\!\!\perp S_2$  and  $E_2 \perp\!\!\!\perp S_1$  (see Drton and Richardson, 2008a, for more details). This amounts to the constraint  $\lambda_{E_1 S_2}^{E_1 S_2} = \lambda_{E_2 S_1}^{E_2 S_1} = 0$ ; the first equality already holds by symmetry, so only one additional constraint is imposed.

This model has a deviance of 8.41 on 7 degrees of freedom, which is not rejected in a likelihood ratio test with the saturated model ( $p = 0.30$ ), and so there is no evidence to reject the separate genes hypothesis. We remark however, that the model with symmetry but no marginal independences has a slightly lower BIC score, and so might be preferred.

The elimination of the 4-way and 3-way interaction parameters for the model from Figure 8(b) with symmetry results in deviances of 11.63 on 8 d.f. and 16.69 on 10 d.f. respectively, both of which also represent reasonable fits; the latter of these has just 5 free parameters.

### 6.3 Netherlands Kinship Data

The Netherlands Kinship Panel Survey (NKPS) is an ongoing study which collects longitudinal information on several thousand Dutch individuals and their families (Dykstra et al., 2005, 2007). One question asked of both the primary respondents (*anchors*) and their partners is “How is your health in general?”, with possible responses of ‘excellent’, ‘good’, ‘good nor poor’, ‘poor’ and ‘very poor’. We combined ‘good nor poor’, ‘poor’ and ‘very poor’ into one category to avoid small counts.

Two waves of data are currently available, from 2002–04 and 2006–07. We only considered anchors who had the same partner in both waves, and such that both the individual and the partner answered the health question in both waves. Let  $A_i$  and  $P_i$  denote the response of the anchor and partner respectively for wave  $i \in \{1, 2\}$ . In total there are  $n = 2,318$  data points, classified into a  $3 \times 3 \times 3 \times 3$  table.

We begin with the complete graph in Figure 9. One plausible model would be that anchors and their partners are exchangeable. Since the graph is symmetrical in this respect, so is the ingenuous parametrization, and enforcing symmetry amounts merely to a set of 36 linear constraints; for example:

$$\lambda_{A_2 P_2}^{A_1 P_1 A_2 P_2}(1, 0) = \lambda_{A_2 P_2}^{A_1 P_1 A_2 P_2}(0, 1).$$

This model has a deviance of 89.98, which when compared to the tail of a  $\chi_{36}^2$  distribution gives  $p = 1.6 \times 10^{-6}$ ; thus the symmetry model is a poor fit to the data, and is rejected. The lack of exchangeability is probably due to selection bias in the sampling of the anchors, as well as the different ways in which the anchors and their partners were asked the question:

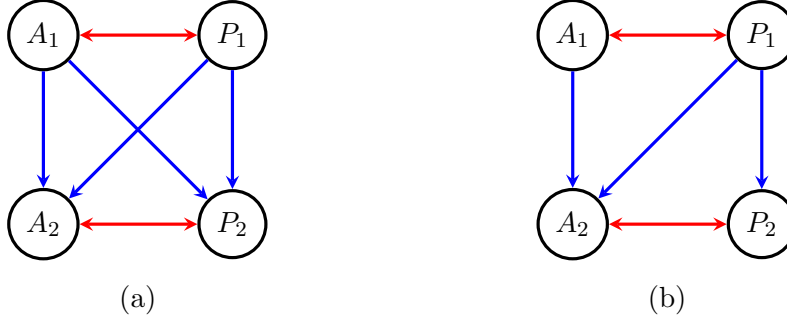


Figure 9: Graphs for the NKPS data; responses of **A**nchor and **P**artner regarding their assessment of health; subscripts indicate time. (a) a complete graph; (b) a subgraph which implies  $P_2 \perp\!\!\!\perp A_1 \mid P_1$ .

anchors were asked about their health as part of a face-to-face interview, whereas the partners were only asked to complete a survey. See Siemiatycki (1979) for an analysis of differences resulting from survey mode.

If instead we remove the edge  $A_1 \rightarrow P_2$  and fit the graph in Figure 9(b), we obtain an explanation of the data which is not rejected at the 5% level (deviance 19.09 on 12 degrees of freedom,  $p = 0.086$ ); this model corresponds to the conditional independence  $P_2 \perp\!\!\!\perp A_1 \mid P_1$ . This graph is the only subgraph of the complete graph in Figure 9(a) which leads to a good fit; in particular the model created by removing the edge  $P_1 \rightarrow A_2$  is strongly rejected, which is one manifestation of the asymmetry between individuals and their partners.

Note that we could also have obtained the independence  $P_2 \perp\!\!\!\perp A_1 \mid P_1$ , for instance, by using a DAG with topological ordering  $P_1, A_1, P_2, A_2$ , but the resulting parametrization would have made it much more difficult to enforce the symmetry constraint tested above.

## 6.4 Example: Trust Data

Drton and Richardson (2008a) examine responses to seven questions relating to trust and social institutions, taken from the US General Social Survey between 1975 and 1994. Briefly, the seven questions were:

**Trust.** Can most people be trusted?

**Helpful.** Do you think most people are usually helpful?

**MemUn, MemCh.** Are you a member of a labour union / church?

**ConLegis, ConClerg, ConBus.** Do you have confidence in congress / organized religion / business?

In that paper, the model given by the graph in Figure 10 is shown to adequately explain the data, having a deviance of 32.67 on 26 degrees of freedom, when compared with the

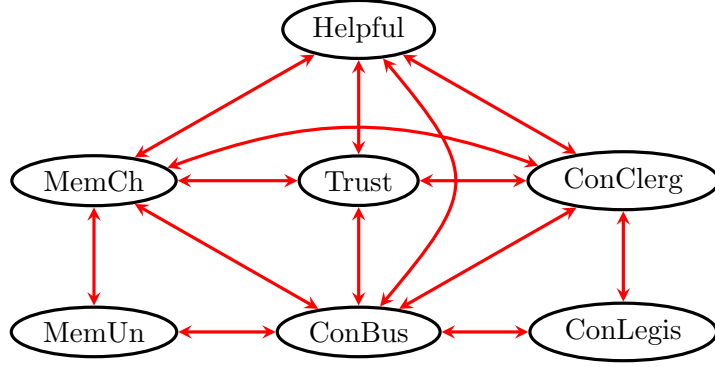


Figure 10: Markov model for trust data given in Drton and Richardson (2008a).

saturated model. The authors also provide an undirected graphical model which has one more edge than the graph in Figure 10, and yet has 62 fewer parameters. It too gives a good fit to the data, having a deviance of 87.62 on 88 degrees of freedom. Both graphs were chosen by backwards stepwise selection methods; see Drton and Richardson (2008a) for details.

For practical and theoretical reasons, the bidirected model may be preferred to the undirected one, even though the latter appears to be much more parsimonious. One may consider the dependence between the responses given to a questionnaire to be manifestations of unmeasured characteristics of the respondent, such as their political beliefs. Such a system can be well represented by a bidirected graph, through its marginal independence structure and connection to latent variable models, but not necessarily by an undirected one, which induces conditional independences. Note that, since models defined by undirected and bidirected graphs are not nested, there is no *a priori* reason to expect the two methods to give a similar graphical structure.

The greater parsimony of the undirected model (when defined purely by conditional independences) is due to its hierarchical nature: if we remove an edge between two vertices  $a$  and  $b$ , then this corresponds to requiring that  $\lambda_A^V = 0$  for every effect  $A$  containing both  $a$  and  $b$ . Removing that edge in a bidirected model may correspond merely to setting  $\lambda_{ab}^{ab} = 0$  and nothing else, depending upon the other edges present. Using the ingenuous parametrization, it is easy to constrain additional higher order terms to be zero to obtain sub-models of the set of distributions obeying the global Markov property.

Starting with the model in Figure 10 and fixing the 4-, 5-, 6- and 7-way interaction terms to be zero increases the deviance to 84.18 on 81 degrees of freedom; none of the 4-way interaction parameters was found to be significant on its own. Furthermore, removing 21 of the remaining 25 three-way interaction terms increases the deviance to 111.48 on 102 degrees of freedom; using an asymptotic  $\chi^2$  approximation gives a p-value of 0.245, so this model is not contradicted by the data. The only parameters retained are the one-dimensional marginal probabilities, the two-way interactions corresponding to edges



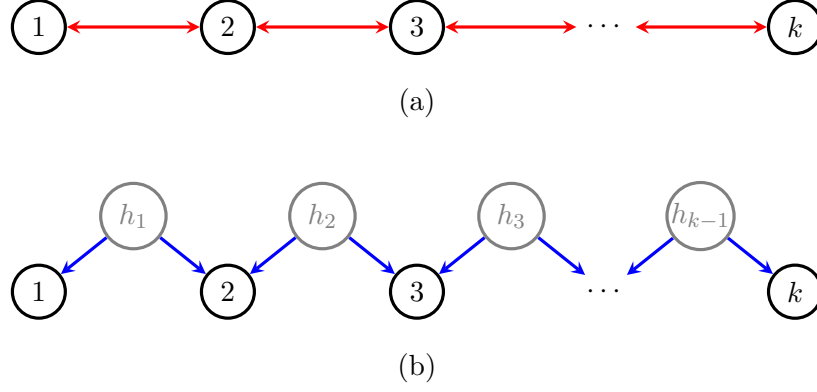


Figure 11: (a) A bidirected  $k$ -chain and (b) a DAG with latent variables  $(h_1, \dots, h_{k-1})$  generating the same observable conditional independence structure.

in Figure 10, and the following three-way interactions:

MemUn, ConClerg, ConBus

Helpful, MemUn, MemCh

Trust, ConLegis, ConBus

MemCh, ConClerg, ConBus.

This model retains the marginal independence structure of Drton and Richardson’s model, but provides a good fit with only 25 parameters, rather than the original 101.

A similar analysis, for different data, is performed by Lupporelli et al. (2009, page 573); again they find an undirected graphical model to be much more parsimonious than any bidirected one, but obtain comparable fits by removing statistically insignificant higher-order parameters.

## 6.5 Simulated Data

We saw in the earlier examples that we were often able to remove higher order interaction parameters without compromising the goodness of fit. Here we explore this phenomena further via simulations.

Consider the DAG with latent variables shown in Figure 11(b); over the observed variables, the conditional independences which hold are exactly those given by the bidirected chain in Figure 11(a).

We randomly generated 1,000 distributions from this DAG model with  $k = 6$ , where each latent variable was given three states, and each observed variable two. The probability of each observed variable being zero, conditional on each state of its parents, was an independent uniform random draw on  $(0, 1)$ ; latent states were fixed to occur with equal probability. For each distribution, a sample size of 10,000 was drawn, and the bidirected chain model was fitted to it by maximum likelihood estimation. For each of the 1,000 data sets, we then measured the increase in deviance associated with removing higher order parameters

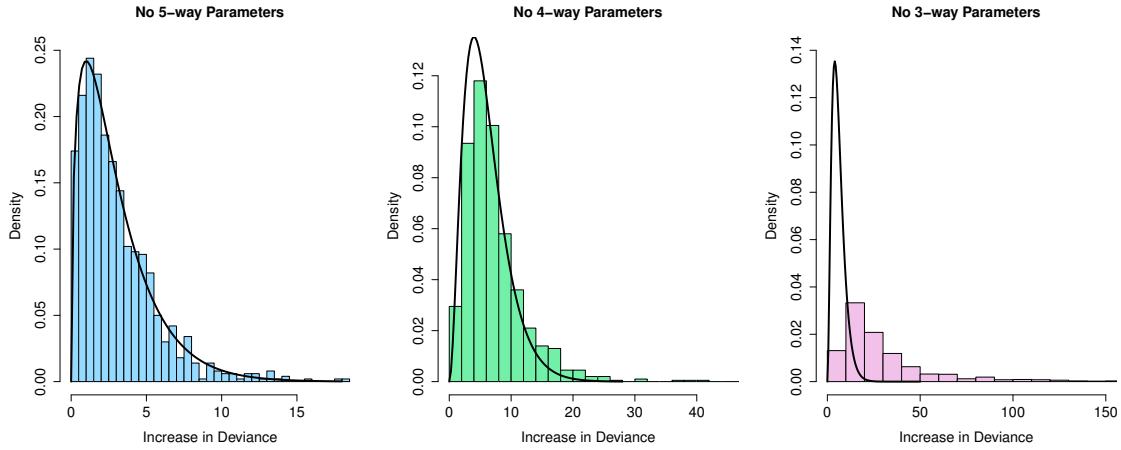


Figure 12: Histograms showing the increase in deviance caused by setting to zero (a) the 5- and 6-way interaction parameters; (b) the 4-, 5- and 6-way interaction parameters; (c) the 3-, 4-, 5- and 6-way interaction parameters. Plots are based on 1,000 datasets, each of size 10,000, generated from the DAG in Figure 11(b). The plotted densities are  $\chi^2$  with 3, 6 and 10 degrees of freedom respectively.

The histogram in Figure 12(a) demonstrates that the deviance increase from setting the 5- and 6-way interaction parameters to zero (a total of three parameters) was not distinguishable from that which would be observed under the null hypothesis that these parameters are zero. The deviance increase from setting the 4-, 5- and 6-way interactions to zero appeared to have only a slightly heavier tail than the associated  $\chi^2$ -distribution, as suggested by the outliers in Figure 12(b). Removing the 3-way interactions in addition to this caused a dramatic increase in the deviance, as may be observed from the heavy tail of the histogram in Figure 12(c). This illustrates that the ingenuous parametrization can be used to produce more parsimonious model descriptions than would be possible using Richardson’s parameters.

Note that under the process which generated these models, each of these interaction parameters was non-zero almost surely. As the sample size increases the power of a likelihood ratio test for a fixed distribution tends to one, so it must be the case that a simulation such as the above would, for large enough data sets, show significant deviation from the associated  $\chi^2$  distributions. However, even at a fairly large sample size of 10,000, a limited effect was observed in Figures 12(a) and (b), and the examples above with real data suggest that higher order interactions are often not particularly useful in practice for describing data.

## 7 Proofs

### 7.1 Proof of Lemma 2.9

*Proof of Lemma 2.9.* Using the independence, we have

$$p_{ABC}(x_{ABC}) = p_{AC}(x_{AC}) \cdot p_{B|C}(x_B | x_C).$$

Thus applying Lemma 2.3,

$$\lambda_{AD}^{ABC}(x_{AD}) = \frac{1}{|\mathfrak{X}_{ABC}|} \sum_{y_{ABC} \in \mathfrak{X}_{ABC}} (\log p_{AC}(y_{AC}) + \log p_{B|C}(y_B | y_C)) \prod_{v \in A \cup D} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1).$$

We can split this sum into terms involving  $p_{AC}(y_{AC})$  and those involving  $p_{B|C}(y_B | y_C)$ .

For the first of these,

$$\begin{aligned} & \frac{1}{|\mathfrak{X}_{ABC}|} \sum_{y_{ABC} \in \mathfrak{X}_{ABC}} \log p_{AC}(y_{AC}) \prod_{v \in A \cup D} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1) \\ &= \frac{1}{|\mathfrak{X}_{AC}| \cdot |\mathfrak{X}_B|} \sum_{y_B \in \mathfrak{X}_B} \sum_{y_{AC} \in \mathfrak{X}_{AC}} \log p_{AC}(y_{AC}) \prod_{v \in A \cup D} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1) \\ &= \frac{1}{|\mathfrak{X}_{AC}|} \sum_{y_{AC} \in \mathfrak{X}_{AC}} \log p_{AC}(y_{AC}) \prod_{v \in A \cup D} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1) \\ &= \lambda_{AD}^{AC}(x_{AC}), \end{aligned}$$

because the summand has no dependence on  $y_B$ . For the latter,

$$\begin{aligned} & \frac{1}{|\mathfrak{X}_{ABC}|} \sum_{y_{ABC} \in \mathfrak{X}_{ABC}} \log p_{B|C}(y_B | y_C) \prod_{v \in A \cup D} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1) \\ &= \frac{1}{|\mathfrak{X}_{ABC}|} \sum_{y_{BC} \in \mathfrak{X}_{BC}} \log p_{B|C}(y_B | y_C) \sum_{y_A \in \mathfrak{X}_A} \prod_{v \in A \cup D} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1). \end{aligned}$$

Now for any  $w \in A$ , the inner part of this term is

$$\begin{aligned} & \sum_{y_A \in \mathfrak{X}_A} \prod_{v \in A \cup D} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1) \\ &= \sum_{y_{A \setminus \{w\}}} \sum_{y_w} \prod_{v \in A \cup D} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1) \\ &= \sum_{y_{A \setminus \{w\}}} \prod_{v \in (A \cup D) \setminus \{w\}} (|\mathfrak{X}_v| \mathbb{I}_{\{x_v = y_v\}} - 1) \sum_{y_w \in \mathfrak{X}_w} (|\mathfrak{X}_w| \mathbb{I}_{\{x_w = y_w\}} - 1) \\ &= 0, \end{aligned}$$

because the innermost summand is  $|\mathfrak{X}_w| - 1$  for precisely one value of  $y_w$ , and  $-1$  for the other  $|\mathfrak{X}_w| - 1$  values. This shows that the whole term is zero, and gives the result.  $\square$

### 7.2 Proof of Lemma 4.4

We first need the following result.

**Lemma 7.1.** For  $L \subseteq M \subseteq V$  with  $N \equiv M \setminus L$ , define

$$\kappa_{L|N}(x_L | x_N) \equiv \sum_{L \subseteq A \subseteq M} \lambda_A^M(x_A).$$

Then

$$\kappa_{L|N}(x_L | x_N) = \frac{1}{|\mathfrak{X}_L|} \sum_{\substack{y_M \in \mathfrak{X}_M \\ y_N = x_N}} \log p(y_M) \prod_{v \in L} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1).$$

*Proof.* Applying Lemma 2.3, we have

$$\begin{aligned} \kappa_{L|N}(x_L | x_N) &= \sum_{L \subseteq A \subseteq M} \frac{1}{|\mathfrak{X}_M|} \sum_{y_M \in \mathfrak{X}_M} \log p_M(y_M) \prod_{v \in A} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) \\ &= \frac{1}{|\mathfrak{X}_M|} \sum_{y_M \in \mathfrak{X}_M} \log p_M(y_M) \sum_{L \subseteq A \subseteq M} \prod_{v \in A} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) \\ &= \frac{1}{|\mathfrak{X}_M|} \sum_{y_M \in \mathfrak{X}_M} \log p_M(y_M) \sum_{L \subseteq A \subseteq M} \prod_{v \in L} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) \prod_{v \in A \setminus L} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) \\ &= \frac{1}{|\mathfrak{X}_M|} \sum_{y_M \in \mathfrak{X}_M} \log p_M(y_M) \prod_{v \in L} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) \sum_{B \subseteq N} \prod_{v \in B} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1). \end{aligned}$$

Now, consider the value of the inner sum, for a fixed  $y_M$ . In the case that there is some  $w \in N$  with  $x_w \neq y_w$ , then

$$\begin{aligned} \sum_{B \subseteq N} \prod_{v \in B} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) &= \sum_{B \subseteq N \setminus \{w\}} \left[ \prod_{v \in B} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) + \prod_{v \in B \cup \{w\}} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) \right] \\ &= \sum_{B \subseteq N \setminus \{w\}} \left[ \prod_{v \in B} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) - \prod_{v \in B} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) \right] \\ &= 0. \end{aligned}$$

Alternatively, if  $x_N = y_N$ , then

$$\begin{aligned} \sum_{B \subseteq N} \prod_{v \in B} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1) &= \sum_{B \subseteq N} \prod_{v \in B} (|\mathfrak{X}_v| - 1) \\ &= |\mathfrak{X}_N| \end{aligned}$$

by the binomial theorem. Thus

$$\kappa_{L|N}(x_L | x_N) = \frac{1}{|\mathfrak{X}_L|} \sum_{\substack{y_M \in \mathfrak{X}_M \\ y_N = x_N}} \log p(y_M) \prod_{v \in L} (|\mathfrak{X}_v|^{\mathbb{I}_{\{x_v = y_v\}}} - 1),$$

since  $\mathfrak{X}_M = \mathfrak{X}_L \times \mathfrak{X}_N$ . □

*Proof of Lemma 4.4.* Let  $N \equiv M \setminus L$ , and pick some  $x_L \in \tilde{\mathfrak{X}}_L$  and  $x_N \in \mathfrak{X}_N$ ; for  $A \subseteq L$ , let  $\mathbf{1}_A$  be a vector of length  $|L|$  with a 1 in position  $j$  if the  $j$ th element of  $L$  is in  $A$ , and 0 otherwise. Define the local  $|L|$ -way log-linear interaction parameter between  $x_L + \mathbf{1}_L$  and  $x_L$  conditional on  $x_N$  as

$$\sum_{A \subseteq L} (-1)^{|L \setminus A|} \log p_{L|N}(x_L + \mathbf{1}_A | x_N);$$

note that since  $x_L \in \tilde{\mathfrak{X}}_L$ ,  $x_L + \mathbf{1}_A \in \mathfrak{X}_L$ . We will first show that we can construct all these local  $|L|$ -way log-linear interaction parameters using the parameters given in the statement of the lemma. As in Lemma 7.1, let  $\kappa_{L|N}(x_L | x_N) \equiv \sum_{L \subseteq A \subseteq M} \lambda_A^M(x_A)$ , and note that

$$\begin{aligned} & \sum_{A \subseteq L} (-1)^{|L \setminus A|} \kappa_{L|N}(x_L + \mathbf{1}_A | x_N) \\ &= \frac{(-1)^{|L|}}{|\mathfrak{X}_L|} \sum_{y_L \in \mathfrak{X}_L} \log p_M(y_L, x_N) \sum_{A \subseteq L} (-1)^{|A|} \prod_{v \in L} \left( |\mathfrak{X}_v| \mathbb{I}_{\{x_v + \mathbb{I}_{\{v \in A\}} = y_v\}} - 1 \right) \end{aligned}$$

follows directly from Lemma 7.1. Now consider the inner sum; if for some  $w \in L$ ,  $y_w \notin \{x_w, x_w + 1\}$ , then

$$\begin{aligned} & \sum_{A \subseteq L} (-1)^{|A|} \prod_{v \in L} \left( |\mathfrak{X}_v| \mathbb{I}_{\{x_v + \mathbb{I}_{\{v \in A\}} = y_v\}} - 1 \right) \\ &= \sum_{A \subseteq L \setminus \{w\}} (-1)^{|A|} \left[ \prod_{v \in L} \left( |\mathfrak{X}_v| \mathbb{I}_{\{x_v + \mathbb{I}_{\{v \in A\}} = y_v\}} - 1 \right) - \prod_{v \in L} \left( |\mathfrak{X}_v| \mathbb{I}_{\{x_v + \mathbb{I}_{\{v \in A \cup \{w\}\}} = y_v\}} - 1 \right) \right] \\ &= 0, \end{aligned}$$

because the value of the outer indicator function is 0 in both terms when  $v = w$ , while the inner indicator functions are the same for all other  $v$ . Alternatively, if  $y_w \in \{x_w, x_w + 1\}$  for all  $w \in L$ , then define

$$B(A) \equiv \{v \in L \mid x_v + \mathbb{I}_{\{v \in A\}} = y_v\}.$$

The map  $A \mapsto B(A)$  is a one-to-one map from  $\mathcal{P}(L)$ , the power set of  $L$ , to itself, i.e. an automorphism. Note that  $D \equiv B(A) \triangle A = \{v \in L \mid x_v = y_v\}$  is independent of  $A$ . Since

$$|A| + 2|B(A) \setminus A| = |B(A)| + |A \triangle B(A)| = |B(A)| + |D|$$

we can rewrite the sum over subsets as

$$\begin{aligned} & \sum_{A \subseteq L} (-1)^{|A|} \prod_{v \in L} \left( |\mathfrak{X}_v| \mathbb{I}_{\{x_v + \mathbb{I}_{\{v \in A\}} = y_v\}} - 1 \right) \\ &= \sum_{A \subseteq L} (-1)^{|B(A)| + |D|} \prod_{v \in L} \left( |\mathfrak{X}_v| \mathbb{I}_{\{v \in B(A)\}} - 1 \right) \\ &= (-1)^{|D|} \sum_{B \subseteq L} (-1)^{|B|} \prod_{v \in L} \left( |\mathfrak{X}_v| \mathbb{I}_{\{v \in B\}} - 1 \right) \\ &= (-1)^{|D|} (-1)^{|L|} \sum_{B \subseteq L} \prod_{v \in B} (|\mathfrak{X}_v| - 1) \end{aligned}$$

which again using the binomial theorem is

$$= (-1)^{|D|}(-1)^{|L|} \prod_{v \in L} |\mathfrak{X}_v| = (-1)^{|D|}(-1)^{|L|} |\mathfrak{X}_L|.$$

Then, substituting this back into the original expression and noting that the two  $(-1)^{|L|}$  factors cancel out,

$$\begin{aligned} \sum_{A \subseteq L} (-1)^{|L \setminus A|} \kappa_{L|N}(x_L + \mathbf{1}_A | x_N) &= \sum_{D \subseteq L} (-1)^{|D|} \log p_M(x_L + \mathbf{1}_{L \setminus D}, x_N) \\ &= \sum_{D \subseteq L} (-1)^{|D|} [\log p_{L|N}(x_L + \mathbf{1}_{L \setminus D} | x_N) + \log p_N(x_N)] \\ &= \sum_{D \subseteq L} (-1)^{|D|} \log p_{L|N}(x_L + \mathbf{1}_{L \setminus D} | x_N), \end{aligned}$$

where the terms in  $\log p_N(x_N)$  cancel because of the lack of dependence upon  $D$ . This is the (conditional) local  $|L|$ -way log-linear interaction. The collection of all the (conditional) local  $|L|$ -way log-linear interactions together with the (conditional)  $(|L| - 1)$ -dimensional marginal distributions smoothly parametrizes the  $|L|$ -way table (Csiszár, 1975; Rudas, 1998).  $\square$

### 7.3 Proof of Theorem 4.7

We require the following lemma.

**Lemma 7.2.** *Let  $\bar{\mathcal{G}}$  be a head-preserving completion of  $\mathcal{G}$ , and let  $H \in \mathcal{H}(\mathcal{G})$  have tails  $T$  and  $\bar{T}$  in  $\mathcal{G}$  and  $\bar{\mathcal{G}}$  respectively. Then under the global Markov property for  $\mathcal{G}$ ,*

$$H \perp\!\!\!\perp (\bar{T} \setminus T) | T [P].$$

*Proof.* Let  $\pi$  be a path in  $\mathcal{G}$  from some  $h \in H$  to  $t \in \bar{T} \setminus T$ , and assume without loss of generality that  $\pi$  does not intersect  $H$  or  $\bar{T} \setminus T$  other than at its endpoints. By Proposition 3.5, every vertex on  $\pi$  is in  $\text{an}_{\mathcal{G}}(\{h, t\} \cup T) \subseteq \text{an}_{\mathcal{G}}(H \cup \bar{T})$ . Since  $\bar{\mathcal{G}}$  is complete, if  $v \in \text{an}_{\bar{\mathcal{G}}}(H \cup \bar{T})$ , then  $v \in H \cup \bar{T}$ , thus  $H \cup \bar{T}$  is ancestral in  $\bar{\mathcal{G}}$ . By Proposition 3.12,  $H \cup \bar{T}$  is also ancestral in  $\mathcal{G}$ , thus every vertex on  $\pi$  is in  $H \cup \bar{T}$ .

By Proposition 3.8,  $\bar{T} \subseteq \text{an}_{\bar{\mathcal{G}}}(H)$ , so  $H \cup \bar{T} = \text{an}_{\bar{\mathcal{G}}}(H)$ . However, since  $H$  forms a head in  $\bar{\mathcal{G}}$ ,  $H$  is barren in  $\bar{\mathcal{G}}$ . Thus in  $\bar{\mathcal{G}}$ , no proper descendant of a vertex in  $H$  is on  $\pi$ , and by Proposition 3.12 this also holds in  $\mathcal{G}$ .

Now let  $y$  be the first vertex after  $h$  on  $\pi$  that is not in  $T$ . By hypothesis,  $y$  exists since  $t \notin T$ . By construction, any vertices between  $h$  and  $y$  on  $\pi$  are in  $T$ , hence are colliders on  $\pi$  and ancestors of  $H$  in  $\mathcal{G}$  (by Proposition 3.8). Thus  $y \in \text{dis}_{\mathcal{G}}(H) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(H))$ . If  $y \in \text{an}_{\mathcal{G}}(H)$  then  $y \in T$ , which is a contradiction, hence  $y \in \text{dis}_{\mathcal{G}}(H)$  and  $y \notin \text{an}_{\mathcal{G}}(H)$ . As shown earlier,  $y$  is not a descendant of a vertex in  $H$ , so  $H \cup \{y\}$  forms a head in  $\mathcal{G}$ . Since  $\bar{\mathcal{G}}$  is a head-preserving completion, it follows that  $H \cup \{y\}$  also forms a head in  $\bar{\mathcal{G}}$ , and thus  $y \notin \text{an}_{\bar{\mathcal{G}}}(H) = H \cup \bar{T}$ , but this is a contradiction.  $\square$

*Proof of Theorem 4.7.* Let  $(H, \bar{T})$  be a head-tail pair in  $\bar{\mathcal{G}}$ . There are three possibilities for how this pair relates to  $\mathcal{G}$ : if  $(H, \bar{T})$  is also a head-tail pair in  $\mathcal{G}$ , then there is no work to be done; otherwise either (i)  $H$  is not a head in  $\mathcal{G}$ , or (ii)  $H$  is a head in  $\mathcal{G}$  but  $\bar{T}$  is not its tail.

If (i) holds, then we claim that under  $\mathcal{G}$ ,  $\lambda_A^{H\bar{T}} = 0$  for all  $H \subseteq A \subseteq H \cup \bar{T}$ . To see this, first note that  $H$  is a barren set in  $\bar{\mathcal{G}}$ , and since  $H$  is maximally connected, this means that all elements are joined by bidirected edges in  $\bar{\mathcal{G}}$ . Since  $\mathcal{G}$  contains a subset of the edges in  $\bar{\mathcal{G}}$ ,  $H$  is also barren in  $\mathcal{G}$ ; since  $H$  is not a head in  $\mathcal{G}$  this means that  $H = K \cup L$  for disjoint non-empty sets  $K$  and  $L$  with no edges directly connecting them. But this implies that  $K$  and  $L$  are m-separated conditional on  $\bar{T}$ , and thus  $X_K \perp\!\!\!\perp X_L \mid X_{\bar{T}}$  under the Markov property for  $\mathcal{G}$ . Then, by Lemma 2.7, these parameters are all identically zero under  $\mathcal{G}$ .

(ii) implies that  $H$  is head in both  $\mathcal{G}$  and  $\bar{\mathcal{G}}$ , but  $\bar{T} \equiv \text{tail}_{\bar{\mathcal{G}}}(H) \supset \text{tail}_{\mathcal{G}}(H) \equiv T$ . Then  $\lambda_A^{H\bar{T}} = 0$  for all  $H \subseteq A \subseteq H \cup \bar{T}$  such that  $A \cap (\bar{T} \setminus T) \neq \emptyset$ ; this follows from Lemma 7.2 and application of Lemma 2.7.

We have shown that all parameters corresponding to effects not found in  $\mathbb{P}^{\text{ing}}(\mathcal{G})$  are identically zero under  $\mathcal{G}$ . The vanishing of these parameters defines the correct sub-model, but note that some of the margins in  $\mathbb{P}^{\text{ing}}(\bar{\mathcal{G}})$  which we have not yet considered are not the same as those in  $\mathbb{P}^{\text{ing}}(\mathcal{G})$ . These remaining cases are again from (ii), but where  $H \subseteq A \subseteq H \cup T$ ; in this case  $\lambda_A^{H\bar{T}} = \lambda_A^{HT}$  under  $\mathcal{G}$ , again due to Lemma 7.2, this time combined with Lemma 2.9.

Thus we have shown that under  $\mathcal{G}$ , all the ingenuous parameters for  $\bar{\mathcal{G}}$  are either zero or equal to ingenuous parameters for  $\mathcal{G}$ . Combined with Theorem 4.5, this shows that those constraints define the model.  $\square$

## 7.4 Proof of Theorem 5.5

We first prove the following graphical result.

**Lemma 7.3.** *Let  $\mathcal{G}$  be an ADMG containing at least one head of size 3 or more. Then  $\mathcal{G}$  also contains two heads of the form  $\{v_1, v_2\}$  and  $\{v_2, v_3\}$ , where  $\{v_1, v_2, v_3\}$  is barren.*

*Proof.* Suppose not; let  $\mathcal{G}$  be an ADMG which violates this condition, and let  $H$  be a head in  $\mathcal{G}$  of size  $k \geq 3$ . Pick 3 vertices  $\{w_1, w_2, w_3\}$  in  $H$ . By the definition of a head, we can pick a bidirected path  $\pi$ , through  $\text{ang}_{\mathcal{G}}(H)$ , from  $w_1$  to  $w_2$ ; assume that  $\pi$  contains no other element of  $H$ , otherwise shorten the path and redefine  $w_1$  or  $w_2$ . Then create a similar path  $\rho$  from  $w_2$  to  $w_3$ ; again assume that  $\rho$  contains no other element of  $H$ , else shorten the path and redefine  $w_3$ . If  $w_1$  lies on  $\rho$  then we can swap  $w_1$  and  $w_2$  to get the desired result.

According to our assumption that the result is false, at least one of  $\{w_1, w_2\}$  or  $\{w_2, w_3\}$  is not a head; assume the former without loss of generality. This implies that  $\pi$  must pass through at least one vertex  $v$  which is not an ancestor of  $\{w_1, w_2\}$ . If there is more than one such vertex, then choose one which has no distinct descendants on the path  $\pi$ . By the

construction of  $\pi$  we have  $v \in \text{ang}_{\mathcal{G}}(H) \setminus H$ .

Then let  $W$  be the set of vertices on  $\pi$ , and  $H^* \equiv \text{barren}_{\mathcal{G}}(W)$ . Since  $W$  is  $\leftrightarrow$ -connected,  $H^*$  must be a head, and  $\{w_1, w_2, v\} \subseteq H^*$ . Thus we have created a head distinct from  $H$ , of size at least 3, which is contained in the set of ancestors of  $H$ .

The assumption we have made implies that we must be able to repeat this process indefinitely, with each head being contained in the ancestors of the previous head. To see that we never obtain the same head twice, note that there is a non-empty directed path from  $v \in H^*$  to  $H$ ; but  $H$  is contained within the ancestors of any previous heads in the sequence, so if  $H^*$  had appeared before, this would imply that  $H^*$  was not barren.

Then since  $H$  has a finite set of ancestors, the apparently infinite recursion of distinct heads is a contradiction.  $\square$

**Definition 7.4.** Let  $A$  be an ancestral set in an ADMG  $\mathcal{G}$ , and let  $v \in \text{barren}_{\mathcal{G}}(A)$ . The *Markov blanket* for  $v$  in  $A$  is the set

$$\text{mb}(v, A) \equiv \text{pa}_A(\text{dis}_A(v)) \cup (\text{dis}_A(v) \setminus \{v\}).$$

In particular, under the ordered local Markov property for  $\mathcal{G}$ ,

$$v \perp\!\!\!\perp A \setminus (\text{mb}(v, A) \cup \{v\}) \mid \text{mb}(v, A). \quad (3)$$

Note that (3) holds for every  $v$  and ancestral set  $A$  (with  $v \in \text{barren}_{\mathcal{G}}(A)$ ) if and only if the global Markov property for  $\mathcal{G}$  holds (Richardson, 2003).

*Proof of Theorem 5.5.* ( $\Leftarrow$ ). Suppose that  $\mathcal{G}$  contains no heads of size  $\geq 3$ , and let  $1, \dots, n$  be a topological ordering on the vertices of  $\mathcal{G}$ . We will construct a complete, hierarchical and variation independent parametrization of the saturated model, and then show that under the global Markov property for  $\mathcal{G}$  it is equivalent to the ingenuous parametrization.

Let  $\mathbb{M}_i \subseteq \mathbb{M}$  be the margins which involve only the vertices in  $[i] = \{1, \dots, i\}$ . Assume for induction, that  $\mathbb{M}_{i-1}$  includes the set  $[i-1]$ , and these margins and their associated effects are hierarchical, complete and satisfy the ordered decomposability criterion up to this point. The base case for  $i = 1$  is trivial.

Now, let the heads involving  $i$  contained within  $[i]$  be  $H_0 = \{i\}, H_1 = \{j_1, i\}, \dots, H_k = \{j_k, i\}$ , where  $j_1 < \dots < j_k < i$  (possibly with  $k = 0$ ). Call the associated tails  $T_0, \dots, T_k$ . We have

$$\text{barren}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(i)) = \{j_k, i\},$$

since  $\text{barren}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(i))$  is a head, and cannot have size  $\geq 3$ . This also implies that  $(H_k \cup T_k) \setminus \{i\} = \text{mb}(i, [i])$ , where  $\text{mb}(v, A)$  is the Markov blanket of  $v$  in the ancestral set  $A$ .

Now, since the ordering is topological,  $A_k \equiv [i]$  is an ancestral set, and the ordered local Markov property shows that

$$i \perp\!\!\!\perp A_k \setminus (\text{mb}(i, A_k) \cup \{i\}) \mid \text{mb}(i, A_k),$$



so

$$i \perp\!\!\!\perp A_k \setminus (H_k \cup T_k) \mid (H_k \cup T_k) \setminus \{i\}.$$

Then for all  $\{i\} \subseteq C \subseteq A_k$  such that  $C \cap \deg(j_k) \neq \emptyset$ ,

$$\begin{aligned} \lambda_C^{A_k} &= \lambda_C^{H_k \cup T_k} && \text{if } H_k \subseteq C \subseteq H_k \cup T_k \\ \lambda_C^{A_k} &= 0 && \text{otherwise,} \end{aligned}$$

where the first equality follows from the independence and Lemma 2.9, and the second from the above independence and Lemma 2.7.

Now set  $A_{k-1} = A_k \setminus \deg(j_k)$ . Then  $A_{k-1}$  is ancestral and contains  $i$ , so applying the ordered local Markov property again gives for any  $\{i\} \subseteq C \subseteq A_{k-1}$  such that  $C \cap \deg(j_{k-1}) \neq \emptyset$ ,

$$\begin{aligned} \lambda_C^{A_{k-1}} &= \lambda_C^{H_{k-1} \cup T_{k-1}} && \text{if } H_{k-1} \subseteq C \subseteq H_{k-1} \cup T_{k-1} \\ \lambda_C^{A_{k-1}} &= 0 && \text{otherwise.} \end{aligned}$$

Continuing this approach gives exactly one parameter for each subset  $C$  of  $[i]$  containing  $i$  and some descendant of any of  $j_1, \dots, j_k$ . Lastly let  $A_0 = A_1 \setminus \deg(j_1)$ . Then for  $\{i\} \subseteq C \subseteq A_0$ ,

$$\begin{aligned} \lambda_C^{A_0} &= \lambda_C^{H_0 \cup T_0} && \text{if } \{i\} \subseteq C \subseteq \{i\} \cup T_0 \\ \lambda_C^{A_0} &= 0 && \text{otherwise.} \end{aligned}$$

Now, add the margins  $A_0 \subset \dots \subset A_k = [i]$ ; since these all contain  $\{i\}$ , they are not a subset of any existing margin. Further, each set  $C$  we associate with  $A_l$  contains a vertex which is not in  $A_{l-1}$ . Thus the addition of these margins and their associated effects keeps our parametrization complete and hierarchical. Setting  $\mathbb{M}_i = \mathbb{M}_{i-1} \cup \{A_0, \dots, A_k\}$ , then there are at most two maximal subsets out of the margins up to  $A_l$  (being  $[i-1]$  and  $A_l$ ); thus  $\mathbb{M}_i$  is clearly also ordered decomposable, and so the parameters are variation independent.

Furthermore we have shown that under the global Markov property for  $\mathcal{G}$ , these parameters are equal to the ingenuous parameters or are identically zero. Thus the ingenuous parameters must also be variation independent.

( $\Rightarrow$ ). Our construction will assume the random variables are binary; the general case is a trivial but tedious extension. Suppose that  $\mathcal{G}$  has a head of size  $\geq 3$ , and assume for a contradiction that its ingenuous parametrization is variation independent. Then by Lemma 7.3, there exist two heads  $H_1 = \{v_1, v_2\}$  and  $H_2 = \{v_2, v_3\}$  such that  $\{v_1, v_2, v_3\}$  is barren. Let  $H_3 \equiv \{v_3, v_1\}$  noting that this set may or may not be a head.

Also let  $T_i = \text{tail}_{\mathcal{G}}(H_i)$ , where if  $H_3$  is not a head, this set is taken to be the tail of  $H_3$  if there were a bidirected arrow between  $v_1$  and  $v_3$ . Further let  $A = \text{ang}(H)$ .

Now choose  $\lambda_{C_i}^{B_i} = 0$ , where  $B_i = \{v_i\} \cup \text{tail}_{\mathcal{G}}(v_i)$  and  $\{v_i\} \subseteq C_i \subseteq B_i$ ; this sets every  $v_i$  to be uniform on  $\{0, 1\}$  for each instantiation of its tail.

Similarly, by choosing  $\lambda_{C_1}^{H_1 \cup T_1}(0)$  to be large and positive for each  $H_1 \subseteq C_1 \subseteq H_1 \cup T_1$ , we can force  $v_1$  and  $v_2$  to be arbitrarily highly correlated conditional on  $T_1$ , and therefore conditional on  $A$ . We can do the same for  $v_2$  and  $v_3$ , so for any  $0 < \epsilon < \frac{1}{2}$ :

	$v_1$			$v_2$	
		0      1			0      1
$v_2$	0	$\frac{1}{2} - \epsilon$ $\epsilon$	$v_3$	0	$\frac{1}{2} - \epsilon$ $\epsilon$
	1	$\epsilon$ $\frac{1}{2} - \epsilon$		1	$\epsilon$ $\frac{1}{2} - \epsilon$

where these tables are understood to show the two-way marginal distributions conditional on each instantiation  $x_A$  of  $A$ .

But now either  $\lambda_{C_3}^{H_3 \cup T_3} = 0$  by design (because  $H_3$  is not a head, and  $v_1$  and  $v_3$  are independent conditional on their ‘tail’), or we can choose this to be the case by the assumption of variation independence. This implies that  $v_1$  and  $v_3$  are independent conditional on  $A$ . Thus

$$\begin{aligned}
\frac{1}{4} &= P(v_1 = 1, v_3 = 0 \mid A = x_A) \\
&= P(v_1 = 1, v_2 = 0, v_3 = 0 \mid A = x_A) + P(v_1 = 1, v_2 = 1, v_3 = 0 \mid A = x_A) \\
&< P(v_1 = 1, v_2 = 0 \mid A = x_A) + P(v_2 = 1, v_3 = 0 \mid A = x_A) \\
&= 2\epsilon,
\end{aligned}$$

which is a contradiction if  $\epsilon < \frac{1}{8}$ . Thus the parameters are variation dependent.  $\square$

## Acknowledgements

This research was supported by the U.S. National Science Foundation grant CNS-0855230 and U.S. National Institutes of Health grant R01 AI032475. The Netherlands Kinship Panel Study is funded by grant 480-10-009 from the Major Investments Fund of the Netherlands Organisation for Scientific Research (NWO), and by the Netherlands Interdisciplinary Demographic Institute (NIDI), Utrecht University, the University of Amsterdam and Tilburg University. We thank McDonald, Hiu and Tierney for giving us permission to use their flu vaccine data.

Our thanks go to Tamás Rudas for helpful discussions, and to Antonio Forcina for discussions and the use of his computer programmes. Finally we thank two anonymous referees and an associate editor for their thorough reading of an earlier draft, and very useful suggestions.

## References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

- R. A. Ali, T. S. Richardson, P. Spirtes, and J. Zhang. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 10–17, 2005.
- W. P. Bergsma and T. Rudas. Marginal models for categorical data. *Ann. Stat.*, 30(1): 140–159, 2002.
- I. Csiszár.  $i$ -divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.
- J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear models for contingency tables. *Ann. Statist.*, 8:522–539, 1980.
- A. P. Dawid. Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B*, 41:1–31, 1979.
- M. Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009.
- M. Drton and T. S. Richardson. Binary models for marginal independence. *J. Roy. Statist. Soc. Ser. B*, 70(2):287–309, 2008a.
- M. Drton and T. S. Richardson. Graphical methods for efficient likelihood inference in Gaussian covariance models. *J. Mach. Learn. Res.*, 9:893–914, 2008b.
- P. A. Dykstra, M. Kalmijn, T. C. M. Knijn, A. E. Komter, A. C. Liefbroer, and C. H. Mulder. Codebook of the Netherlands Kinship Panel Study, a multi-actor, multi-method panel study on solidarity in family relationships. Wave 1. *NKPS Working Paper No. 4*, 2005.
- P. A. Dykstra, M. Kalmijn, T. C. M. Knijn, A. E. Komter, A. C. Liefbroer, and C. H. Mulder. Codebook of the Netherlands Kinship Panel Study, a multi-actor, multi-method panel study on solidarity in family relationships. Wave 2. *NKPS Working Paper No. 6*, 2007.
- A. Ekholm, J. Jokinen, J.W. McDonald, and P.W.F. Smith. A latent class model for bivariate binary responses from twins. *J. Roy. Statist. Soc. Ser. C*, 61(3), 2012.
- R. J. Evans and A. Forcina. Two algorithms for fitting constrained marginal models. Arxiv preprint arXiv:1110.2894, 2011.
- R. J. Evans and T. S. Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th conference on Uncertainty in Artificial Intelligence*, 2010.
- A. Forcina, M. Lupporelli, and G. M. Marchetti. Marginal parameterizations of discrete models defined by a set of conditional independencies. *Journal of Multivariate Analysis*, 101:2519–2527, 2010.

- G. F. V. Glonek and P. McCullagh. Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B*, 57(3):533–546, 1995.
- A. J. Hakim, L. F. Cherkas, T. D. Spector, and A. J. MacGregor. Genetic associations between frozen shoulder and tennis elbow: a female twin study. *Rheumatology*, 42(6):739–742, 2003.
- G. Kauermann. A note on multivariate logistic models for contingency tables. *Austral. J. Statist.*, 39(3):261–276, 1997.
- S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, UK, 1996.
- M. Lupparelli, G. M. Marchetti, and W. P. Bergsma. Parameterizations and fitting of bi-directed graph models to categorical data. *Scand. J. Statist.*, 36:559–576, 2009.
- G. M. Marchetti and M. Lupparelli. Chain graph models of multivariate regression type for categorical data. *Bernoulli*, 17(3):827–844, 2011.
- C. J. McDonald, S. L. Hui, and W. M. Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD computing*, 9(5):304, 1992.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.
- T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Statist.*, 30(1):145–157, 2003.
- T. S. Richardson. A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the 25th conference on Uncertainty in Artificial Intelligence*, 2009.
- T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann. Statist.*, 30:962–1030, 2002.
- T. Rudas. *Odds ratios in the analysis of contingency tables*. Sage Publications, Inc, 1998.
- T. Rudas, W. P. Bergsma, and R. Németh. Parameterization and estimation of path models for categorical data. In *Proceedings in Computational Statistics, 17th Symposium*, pages 383–394. Physica-Verlag HD, 2006.
- T. Rudas, W. P. Bergsma, and R. Németh. Marginal log-linear parameterization of conditional independence models. *Biometrika*, 97:1006–1012, 2010.
- J. Siemiatycki. A comparison of mail, telephone, and home interview strategies for household health surveys. *American Journal of Public Health*, 69:238–245, 1979.
- N. Wermuth. Probability distributions with summary graph structure. *Bernoulli*, 17(3):845–879, 2011.
- J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley, 1990.